12-2021

On Predicting Omnidirectional Honey Bee Traffic Using Weather and Electromagnetic Radiation

Daniel G. Hornberger
Utah State University

ON PREDICTING OMNIDIRECTIONAL HONEY BEE TRAFFIC USING WEATHER AND ELECTROMAGNETIC RADIATION

by

Daniel G. Hornberger

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

Vladimir A. Kulyukin, Ph.D. Major Professor	Nicholas S. Flann, Ph.D. Committee Member
John Edwards, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2021

ABSTRACT

On Predicting Omnidirectional Honey Bee Traffic Using Weather and Electromagnetic Radiation

by

Daniel G. Hornberger, Master of Science

Utah State University, 2021

Major Professor: Vladimir A. Kulyukin, Ph.D.
Department: Computer Science

Honey bee populations have declined significantly since 1961. While some causes of this decline are known, others are not. By utilizing electronic bee hive monitoring (EBM) systems, apiarists and researchers have an added resource in determining the causes of these declines so that the issues can be remedied. For nearly five months (May through October) during the 2020 honey bee foraging season in Logan, Utah, USA, we collected on-site weather and electromagnetic radiation (EMR) readings and videos of the hive entrances of six bee hives every 15 minutes. Each video was processed to estimate the number of bee motions, and the bee motion counts were paired with the weather and EMR data. We show that by using a random forest regressor trained on select weather variables coupled with omnidirectional bee motion counts obtained by a video-based EBM system, bee motion at the hive entrance can be estimated with R^2 scores ranging from 0.547 every 15 minutes up to 0.886 every 12 hours. We suggest that these predictions can be used to trigger remote alerts for bee keepers when the observed behavior significantly differs from the predicted values. We also propose that the nine-hour relative humidity trend is a new major predictor variable in addition to solar radiation and temperature. Additionally, we show that nearly 19% of the variance in the motion of bees at the hive entrance can be explained by the

variance in select EMR variables. In conjunction with these findings, we are contributing the design and associated software for our Weather and EMR Sensing Station, along with 12 weather, EMR, and motion count data files to the public for further exploration. We also provide an analysis of the data for one of the hives, and discuss various correlations, relations, and observations pertaining to variables that may be influencing bee behavior.

(119 pages)

PUBLIC ABSTRACT

On Predicting Omnidirectional Honey Bee Traffic Using Weather and Electromagnetic Radiation

Daniel G. Hornberger

Honey bees are responsible for pollinating many important crops in the United States. However, honey bee populations have declined significantly since 1961. While some causes of this decline are known, others are not. By utilizing electronic bee hive monitoring (EBM) systems, bee keepers and researchers have an added resource in determining the causes of these declines so that the issues can be remedied. For nearly five months (May through October) during the 2020 honey bee foraging season in Logan, Utah, USA, we collected on-site weather and electromagnetic radiation (EMR) readings and videos of the hive entrances of six bee hives every 15 minutes. Each video was processed to estimate the number of bee motions, and the bee motion counts were paired with the weather and EMR data. We show that an algorithm consisting of many decision trees in concert (called a random forest regressor) can be used to accurately predict bee motion counts from one moment to another. By using specific weather variables coupled with omnidirectional bee motion counts, we demonstrate that 54.7% of bee motion at the hive entrance can be predicted accurately with data records spaced 15 minutes apart, and 88.6% of the bee motion can be predicted accurately for those 15 minute measurements averaged over 12 hours. We suggest that these predictions can be used to trigger remote alerts for bee keepers when the observed behavior significantly differs from the predicted values. We also propose that the nine-hour relative humidity trend is a major predictor variable in addition to solar radiation and temperature. Additionally, we show that nearly 19% of the changes in bee motion at the hive entrance can be explained by the changes in select EMR variables. In conjunction with these findings, we contribute the design and software for our Weather and EMR Sensing Station, along

with 12 curated data files to the public for further exploration. We also discuss various correlations, relations, and observations pertaining to variables that may be influencing bee behavior. The associated expenses of this research was approximately 1,500 US dollars.

To my incredible wife Anna who inspires me to be better every day, and my parents Michael and Jenet for instilling in me the love of learning.

ACKNOWLEDGMENTS

I would like to thank the many people who have helped me throughout the completion of my master's degree at Utah State University. I have grown and learned more than I believed I could, and I'm grateful for all those who made this possible.

I would like to thank my major professor Dr. Vladimir Kulyukin for his kind support over the last couple years. He has been a wealth of experience and knowledge, and his sincere desire to find truth has been inspiring. His feedback has been invaluable, and his questions, guidance, and ideas have greatly improved this paper. Additionally, he performed countless hours of work to process each of the videos collected to obtain the bee motion count data we used in this research. Most of all, I would like to thank him for his sincere interest in the well-being of me and my family. His character and kindness is an example to all who work with him, and I'm grateful to be able to call him my friend.

I'm also grateful to the other members of my committee, Dr. Nicholas Flann and Dr. John Edwards, for their valuable feedback and comments.

I would also like to thank my wonderful wife Anna Hornberger for her ideas, constant encouragement, detailed reviews, and for helping me see outside the box. She also designed the physical structure of the Weather and EMR Sensing station, and helped put together the part orders. Her tireless service and her positive belief in me throughout these last few busy years is what has allowed me to complete this research.

I'm also grateful for the extremely helpful statistical advice and direction given to me by Tyler (Pablo) Olsen and Lacy Imes. They helped me start off on the right foot from the beginning.

I would also like to acknowledge the helpful reviews, revisions, questions, and ideas from Michael, Jenet, and Thomas Hornberger. These helped me improve the overall quality of the paper, and helped me better convey my ideas in a way that they could make sense to other readers.

Additionally, I would like to thank my three Graduate Program Coordinators, Genie Hanson, Kaitlyn Fjeldsted, and Caitlin Thaxton who were able to answer my questions and help me meet each of the degree requirements on time. I would also like to thank Cora Price for her help in ordering the many parts needed to build the two Weather and EMR Sensing stations.

We would like to thank all the Kickstarter backers and, especially, our BeePi Angel Backers (in alphabetical order): Prakhar Amlathe, Ashwani Chahal, Trevor Landeen, Felipe Queiroz, Dinis Quelhas, Tharun Tej Tammineni, and Tanwir Zaman. All computer hardware and some hive woodenware and bee packages were procured with the resources from our Kickstarer Open Science crowdfunders in 2017 and 2019. We would also like to express our gratitude to Richard Waggstaff, Craig Huntzinger, and Richard Mueller for letting us use their property for longitudinal electronic beehive monitoring experiences.

And lastly, I would like to thank the Space Dynamics Lab along with my co-workers and supervisors for supporting me in gaining additional education. Their willingness to work with my class and meeting schedule has been greatly appreciated.

Daniel G. Hornberger

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ADC	Analog to Digital Converter
AICc	Corrected Akaike Information Criterion
CCD	Colony Collapse Disorder
CO_2	Carbon Dioxide
CSV	Comma Separated Values
CART	Classification and Regression Tree
DECT	Digital Enhanced Cordless Telecommunications
EBM	Electronic Beehive Monitoring
EF	Electric Field
ELF	Extremely Low Frequency
EMF	Electromagnetic Field
EMR	Electromagnetic Radiation
GB	Gigabyte
GHz	Gigahertz
GPIO	General-purpose Input/Output
GPLv3	GNU General Public License v3.0
HAT	Hardware Attached on Top
HCC	Health Colony Checklist
KNN	K-Nearest-Neighbors
LAN	Local Area Network
MB	Megabyte
MHz	Megahertz
MPH	Miles per Hour
MSE	Mean Squared Error
NaN	Not a Number
PCA	Partial Component Analysis

PLS	Partial Least Squares
RF	Radio Frequency
RFE	Recursive Feature Elimination
RFECV	Recursive Feature Elimination with Cross-validation
RFID	Radio Frequency Identification
RFR	Random Forest Regressor
RTC	Real Time Clock
SD	Secure Digital
SFS	Sequential Feature Selection
SNR	Signal to Noise Ratio
SSH	Secure Shell
USB	Universal Serial Bus
USDA	United States Department of Agriculture
USU	Utah State University

CHAPTER 1

INTRODUCTION

Honey bees have been useful to humans for thousands of years. Hive products such as honey, pollen, royal jelly, beeswax, propolis, and venom have various nutritional and medicinal uses by people [1]. Not only do the hive products have value, but also the pollination work that bees perform on agricultural crops. Honey bees are responsible for pollinating approximately 33% of the food we consume, and they increase the crop production revenue in the United States by more than 15 billion dollars each year [2].

However, honey bee colonies face many challenges. Honey production has decreased in the United States by more than 30% since 1987 [3]. In the year 2020, about 35% of the bee hives in the United States were lost [4]. Between 1961 and 2007, honey bee colonies decreased by more than 49% in the United States, and more than 26% in Europe [5].

While these declines in bee populations could have many causes, some that are known include mites, parasites, pests, diseases, and pesticides. During the year 2020, as much as 25% to 55% of the colonies in the United States were affected by Varroa mites throughout the year [4]. There are also instances where a lost colony exhibits symptoms that aren't consistent with any known causes of death. These occurrences are known as Colony Collapse Disorder (CCD). According to the U.S. Department of Agriculture (USDA), CCD is accompanied with a sudden loss of the hive's mature bee population with very few bees found around the hive, several frames of healthy, capped brood with low levels of parasitic mites, food reserves that haven't been robbed due to other colonies avoiding the hive, little evidence of wax moth or small hive beetle damage, and a laying queen often present with a small cluster of newly emerged attendants [6]. These confounding symptoms indicate that the hive was relatively healthy before suddenly collapsing. Over seven percent of the colonies in the United States were lost due to CCD in 2020 [4]. Bee keepers and researchers alike still don't know the exact cause of CCD or how to prevent it.

There is speculation that other factors not mentioned above may be adversely affecting bee populations as well. These include things such as global warming [7] [8], environmental pollution [9], reduced genetic variability [5], higher levels of electromagnetic radiation or electric fields [5] [10] [11], and loss of foraging pasture [5].

Honey bees have also been used as biological indicators to measure the health of surrounding environments [12] [13]. Since honey bees regularly collect small samples of soil, air, water, and plants during their foraging activities, scientists can analyze their honey, pollen, and mortality to reveal information about dangerous pollutants in the area. For example, a study was performed shortly after the Chernobyl nuclear disaster where honey and pollen were collected and analyzed at a site nearly 1,000 miles away from the event in 1986, and the researchers found concentrations of harmful fallout deposits [14]. In another study, researchers monitored death rates of honey bees near farmland in Italy where certain fungicides were being utilized [12]. As death rates increased, the bees were analyzed and often found to contain chemicals with very high cancer-causing potential [12]. Additionally, another study showed that monitoring bees can be a valuable asset to quantify the changing of a natural environment due to human activity [15].

Since some of the environmental factors affecting bees can and may already be affecting humans, decreasing bee populations raises the flag of warning for us as well. Where many of the elements causing this trend are unknown, it becomes our best interest to discover what they are, and whether they will begin or continue to negatively affect our well-being now and in the future. Regardless of whether these unknown elements are affecting us directly, the consequences of decreasing bee populations will influence us indirectly in many other ways.

One mechanism that can assist bee keepers and researchers in understanding what may be influencing bee population declines is regular hive checks to assess bee hive health. This can allow bee keepers to catch problems early enough to remedy them, as well as discover what stressors may be present. However, apiarists often manage many bee hives that are geographically diffuse. This makes hive checks difficult, time consuming, and

costly. Further, hive checks are intrusive and disruptive to a colony's daily activities. Thus, problems with a hive may go unnoticed until it is too late to save a colony or determine what caused its demise. Fortunately, advances in technology have made Electronic Beehive Monitoring (EBM) a feasible route to solving many of these problems.

EBM is made possible by using various sensors in, on, or around a hive to measure and estimate the current activity of a honey bee colony. EBM approaches vary widely from one to another, and each method has its advantages and disadvantages. Some methods include using infrared trip lines or Radio Frequency Identification (RFID) chip sensors at the hive entrance [16] [17], capturing beehive audio samples to estimate the state of a colony [18], hive scales to monitor the weight of a hive over time [19], machine learning or Digital Particle Image Velocimetry on video footage to estimate directional bee motion [20] [21], and in-hive metrics such as temperature, humidity, and $C0_2$ levels [16] [22].

By making use of one or more of these methods of EBM, it's possible to utilize various algorithms on the collected bee hive monitoring data to make predictions about what the behavior of a colony should be at a given time or season. This can allow alerts to be transmitted to apiarists when colony behavior differs significantly from what the algorithm predicts it should be. By giving greater alerting capability to bee keepers, it may be possible for them and researchers to gain further insight into what influences bee behavior on a daily basis, as well as provide a means whereby unknown causes of bee declines can be discovered.

The goal of this research is to utilize environmental conditions, such as weather and ambient electromagnetic radiation (EMR), paired with data from an EBM system (the BeePi Monitor) to accurately predict omnidirectional bee traffic levels throughout the day. Additionally, we aimed to gain better insight into what environmental factors contribute most to predicting bee motion at the hive entrance, and to determine whether EMR at our selected hive site appeared to significantly contribute to those predictions. And finally, we also aimed to produce data sets and cost effective means whereby other researchers could perform further experiments and collect additional data.

1.1 Related Work

Many researchers have conducted various experiments to determine how different environmental factors affect bee behavior. An early study in 1981 performed by Burrill and Dietz [23] analyzed the effects solar radiation intensity, temperature, relative humidity, and atmospheric pressure on honey bee behavior. By using a photoelectric bee counting and recording device, the researchers were able to observe how these environmental factors correlated with bee traffic.

Burrill and Dietz's results showed that the temperature and solar radiation had the largest influence on bee foraging activity. While outgoing bee traffic positively corresponded with increasing temperatures, they also observed that bee traffic increased with solar radiation levels up to about 0.66 Langleys[1]. Once the solar radiation intensity rose above that threshold, bee traffic levels began to decrease. This finding that temperature and solar radiation have the largest influence on bee foraging activity has since been validated by several other studies [24] [25] [26] [27].

In 2003, Devillers et al. [26] conducted similar experiments while adding additional weather variables in an effort to predict outgoing bee traffic. These additional variables included rainfall and wind and from weather station 6 km away from the hive site. Atmospheric pressure was not used. After collecting 54 days of hourly readings, they performed Principal Component Analysis (PCA) to narrow down the input variables to a smaller subset of uncorrelated ones, and then fitted a Partial Least Squares (PLS) regression model to the data. This process allowed them to achieve R^2 scores for hourly predictions ranging from 0.620 to 0.715, indicating that up to 71.5% of the variance in bee traffic was explained by the variance of their collected input variables. It was noted that their model typically underestimated particularly high activities at the hive entrance.

Devillers et al. also performed analyses on their collected data to gain insight into how each weather variable appeared to influence bee behavior. The research reported that they didn't find any significant co-structures between the humidity, wind, or rain with bee activity at the hive entrance. However, it was mentioned that the wind and rain data enabled the

[1] A Langley is unit of heat transmission defined as 1 thermochemical calorie per square centimeter.

explanation of specific situations. They observed that bees were more active after a short rainy event, and that the relative humidity extremes provided useful information in their models.

In 2014, Polatto et al. [25] performed a study where bee foraging activity was estimated at a location away from the hive entrance by regularly capturing bees at foraging sites and counting them. These bee counts were collected hourly, and weather variables for temperature, humidity, luminosity, and wind speed were also recorded. This study reported finding a negative correlation of relative humidity to honey bee activity. Also, by fitting a linear function to the temperature and luminosity values, they obtained an R^2 score of 0.618, which is similar to the results reported by Devillers et al. [26]. Additionally, they reported the Pearson correlation strength of each collected weather variable with the foraging activity of *Apis mellifera*: 0.617 for temperature, -0.800 for relative humidity, 0.827 for luminosity and 0.793 for wind speed.

Clarke and Robert improved upon the R^2 scores achieved by Devillers et al. [26] and Polatto et al. [25] in 2018 [24]. Their study included a more complete set of weather variables (temperature, solar radiation, atmospheric pressure, humidity, rainfall, wind direction, and wind speed) than the previous studies. Like Devillers et al. [26], they used an electro-optical bee counter to continuously count bee traffic at the hive entrance. Their sampling rate of bee traffic and weather was much narrower at one per minute. They used data that was normalized and combined from one hive during one season and another hive during the next season to train and test various Partial Least Squares predictive models.

When using only temperature and solar radiation as inputs, Clarke and Robert were able to obtain an R^2 score of 0.78. With the addition of atmospheric pressure as an input, they reached an R^2 score of 0.81. The other collected variables where shown to have minimal contributions to the performance of their models. It should be noted that while the data used for this study was at one minute intervals, the data periodicity used for the reported results did not appear to be specified. Since the results were compared directly with those of Devillers et al. [26], it's assumed that the periodicity was the same at one hour.

Clarke and Robert speculate that their results may have been better because they used data from multiple hives across multiple seasons to generate the model. While the results of Devillers et al. are not as good as those by Clarke and Robert, since Deviller's group didn't report the initial correlations of their weather data with bee traffic, the results between the two studies may not be comparable. Clarke and Robert reported Pearson correlations for solar radiation, temperature, and humidity at 0.72, 0.78, and -0.75 respectively. These correlations are very similar to those reported by Polatto et al. [25], which correspond to their R^2 score of 0.618.

Clarke and Robert also estimated the performance of their models at 24 hour and one week temporal resolutions by averaging the predicted and measured egress rate over those time periods and calculating the error. This resulted in an error of 0.3825 for 116 data points. In doing so, they postulate that their model could be applied to different temporal resolutions and provide useful predictions.

In another study, Zacepins et al. [28] utilized a single thermometer placed inside the hive at the top to remotely detect and predict colony swarming events. Temperature readings were logged each minute, and hives were visually monitored by a beekeeper at the hive site. By monitoring the changes in temperature immediately preceding swarming events, they showed that the temperature rose sharply between eight and 20 minutes before take-off. An algorithm was then developed that utilized the temperature readings to accurately predict swarming events. This method proved to be very accurate as tested on a small amount of data, but it only monitored one aspect of hive activity.

A recent study performed by Braga et al. [22] approached hive distress detection and prediction based on the Health Colony Checklist (HCC). Their approach utilized the internal beehive temperature and hive weight along with the external weather variables temperature, dew point, wind direction, wind speed, rainfall, and daylight in combination with weekly apiary inspection results. The weekly apiary inspection results consisted of six different categories of health which were used as the labels for the data. These included the presence of all brood stages, sufficient adult bee populations, the presence of a young laying queen,

sufficient nutrition, the absence of apparent hive stressors, and suitable space in the hive. It should also be noted that at each weekly inspection the internal thermometer was moved to be close to the center of the brood area.

After collecting the data and associating it with the proper health statuses, Braga et al. utilized various supervised machine learning models to predict the hive health. These models included K-Nearest-Neighbors, random forests, and neural networks. By using the random forest model, they were able to reach a hit rate of 98%.

While this method of hive health prediction reached the highest level of accuracy, the labels used are such that the model is structured for longer-term estimations. For instance, a swarming or robbing event likely wouldn't be detected until after the fact when the hive begins to exhibit behavior associated with failed HCC items. Additionally, the hive monitoring process as described wouldn't be able to be completely autonomous because it required the thermometer to be re-positioned on a weekly basis.

Other research has emphasized certain environmental factors that influence bee activity that could also be used to predict bee activity by algorithms coupled with electronically monitored beehives. For instance, in 2015, Alves et al. [29] performed a study very similar to that of Polatto et al. [25] where bees were manually captured at foraging sites to estimate foraging activity and how it correlated with temperature, humidity, and wind speed. Their data exhibited Spearman correlations of -0.691 for relative humidity and 0.531 for temperature. They reported that when the relative humidity rose above 81% there was no foraging activity, but that temperature and humidity accounted for 46.9% of the activity observed. Additionally, they reported that peak foraging was observed when the temperature was high at around 84°F and the relative humidity was low at around 43.6%. Their research also showed that the wind speed didn't appear to significantly affect foraging activity.

Likewise, a study by Jiang et al. in 2016 observed the influence of humidity on bee activity [16]. In their research, they developed a system that could count bees entering and leaving the hive and capture data from various sensors. They reported that humidity inside the hive affected aspects such as egg hatching, sealed brood percentage, and flight

activity. They also indicated that it appeared the optimum humidity range associated with peak bee activity was 60 to 70%, which is higher than reported by Alves et al [29]. These differences could be attributed to a number of factors, such as location or observation methods. Additionally, Jiang et al. noted that low traffic levels were recorded on rainy days.

Mattos et al. [27] in 2018 performed a study where they manually counted worker bees returning to the hive after foraging while recording 12 different weather variables. Along with observing a negative correlation between honey bees and relative humidity, they also stated that there appeared to be a slight positive relationship between pollen collection and wind speed. In addition to the variables the previously mentioned studies used, this one incorporated several novel weather variables. These included the minimum and maximum temperature, the height of the first base of clouds, and the three-hour atmospheric pressure variation. However, after statistically analyzing these along with the more conventional variables, they determined that none of the new variables were strongly correlated to the pollen collection observed, and they were left out of their linear equation used to fit returning bee traffic.

In 2020, a study was performed by Hennessy et al. [30] that explored the effects of wind on honey bee foraging activity. To do this, they set up artificial flowers containing nectar at different spacings, and used electric fans to simulate wind. Honey bees were then observed while foraging at different fan speeds and flower movements. They found that higher wind speeds correlated with longer hesitancy to take off, which in turn prevented the bees from visiting as many flowers. They also found that if the wind speed reached about 5 MPH, foraging activity would stop.

A study in 2016 by Xujiang et al. [31] monitored honey bees using RFID tags to determine if bees work harder before a rainy day. They monitored bee traffic for 34 days and associated traffic levels of each day with the conditions of the following day. In doing this, they discovered that the bees spent more time outside the hive on days preceding rainy days compared to those followed by sunny days. This led them to speculate that honey bees may be able to sense changes in environmental factors such as barometric pressure,

humidity, and temperature to anticipate changes in the weather.

In addition to traditional weather variables affecting bee behavior, research has shown that other anthropogenic factors, such as EMR, influence bee activity as well. Thus, this environmental factor could also potentially be used to predict bee activity. Kimmel et al. [32] performed an experiment in 2007 to assess the effects of EMR on a bee hive by placing a Digital Enhanced Cordless Telecommunications (DECT) phone at the bottom of 16 bee colonies. This device emitted non-ionizing EMR at 1900 MHz with 2.5 mW of power. Some of the hives received full exposure to the device, others received 50% exposure, and the remaining hives were shielded from EMR. They then took some of the bees from each hive 500 meters away, and timed how long it took for them to return to their hives. This revealed that the bees exposed to the EMR generally took longer to return than bees that weren't exposed, showing that EMR adversely affected these bees.

Another study performed by Taye et al. [33] in 2017 analyzed the effects of a cell tower on various bee hives spaced a different distances from the tower. They monitored the RF at each hive while counting the number of worker bees leaving the hive, and the number of worker bees returning to the hive with pollen each minute for 15 days. They found that the return rate of the bees was the smallest for the hive located nearest to the cell tower.

Power lines have also been shown to impact bee behavior. In 1981, Greenberg et al. [11] performed a study where 100 shielded and unshielded bee hives were placed under a 765 kV power line and monitored for several months. They observed that hive weight gain was slowed after two weeks of exposure, and after four weeks queen loss, abnormal queen production, and colony failure was also recorded. If a hive was removed from exposure, they noticed that the hive began gaining weight at a healthy rate again. The researchers concluded that electric fields produced by power lines created environments wherein honey bees could not healthily live.

Another study was performed in 2019 by Shepherd et al. [10] that analyzed the effects of Extremely Low Frequency (ELF) Electromagnetic Frequencies (EMF) generated under power lines on honey bees. Since power lines can generate ELF EMF up from 100 to 1000

micro Teslas, the researchers individually exposed 357 bees to one of those two levels of ELF EMF. After exposing them for a period of time, they performed various tests to estimate their level of aversive learning and aggression. They found that the exposure reduced how well the bees learned, and caused them to exhibit higher aggression levels.

1.2 Our Contributions

Each of the papers referenced in the previous section have various strengths and shortcomings. While most of the papers discussed here used conventional weather variables as predictors, only the work done by Mattos et al. [27] attempted to use a variable that represented a change over time: Atmospheric pressure variance over three hours. While it was found in this study that it didn't correlate significantly with bee activity, perhaps the trend interval amounts or trend representations for other variables could be useful in predicting bee activity. It may be possible that honey bees are responding to *changes* in environmental conditions, rather than whatever the conditions currently are. This may reveal that environmental factors such as humidity, wind speed, or wind direction (in addition to solar radiation and temperature) could be valuable predictors as well when their change over time is represented appropriately.

Also, most of the effective predictive methods used Partial Least Squares regression to create a model to estimate bee activity for given weather inputs. The work by Braga et al. [22] achieved the highest prediction accuracy by using a random forest model. However, this was when only six output classes were being used, and the method was limited to predicting hive health statuses over time. It wouldn't be able to predict events such as swarming or robbing. However, a random forest regressor could be used to predict bee foraging activity throughout the day to allow for more timely predictions, and could result in useful accuracies.

Additionally, while several methods mentioned here used conventional weather metrics to predict bee foraging activity, none of them used EMR as an environmental input to attempt to predict foraging traffic. As discussed in the previous section, several papers have shown obvious effects of EMR on honey bee activity. With the advent of new wireless

technologies, it may be possible that daily fluctuations in ambient EMR is affecting bee activity from one moment to the next. Perhaps ambient EMR levels could be used as inputs in addition to climatic variables to improve the accuracy of bee foraging activity prediction models.

In this thesis, we explore the efficacy of using EMR, conventional weather variables, weather trend variables, and a random forest regressor to predict omnidirectional bee traffic, which in turn can be used to estimate hive health over the short and long term. We show that a new variable, nine-hour relative humidity trend, exhibits better foraging predictive power than temperature. We also demonstrate that this humidity trend variable, along with other conventional and new trend variables, can be used to generate predictions with different accuracies and confidence intervals over various temporal resolutions. And finally, we explore the viability of using ambient EMR at our particular data collection site with our selected EMR sensor to predict bee foraging activity.

We will also discuss our cost effective means of collecting weather and EMR data at the hive site, how we fused the weather and EMR data with the bee motion count data from the BeePi device, and how the data was preprocessed to produce 12 new data sets. We also present an analysis of how each variable correlates with bee activity along with other observations that influenced our model design decisions. We then discuss the process of developing, testing, and interpreting our model to predict omnidirectional honey bee motion at the hive entrance.

It should be noted that many other bee hive monitoring approaches utilize means that require physically altering or placing sensors inside the beehive. However, approaches that require physical hive modifications may decrease the chances that apiarists will adopt such a method. Also, these modifications along with other in-hive sensors can also affect honey bee phenology. And lastly, many professional bee monitoring solutions can be quite costly.

The research efforts described here are a continuation of a larger research effort that adheres to several guiding principles. As stated by Kulyukin and Mukherjee [20], these include ensuring that the EBM hardware, design, and software is 100% replicable, that

the EBM system should be compatible with standard beehive models used by beekeepers worldwide, and that the bee space is preserved such that the deployment of the EBM sensors should not be disruptive to any natural bee colony cycles.

The remainder of this thesis is structured as follows. In Chapter 2, we describe the hardware components, structure of the Weather and EMR Sensing Station, the software, and the variables collected by the station. In Chapter 3, we discuss how we collected our data, how the data from the Weather and EMR Sensing Station and BeePi monitors was fused, how we curated the data, and we enumerate the data files made available to the public. In Chapter 4, we provide an analysis of the data for one of the hives and identify important correlations and trends between important variables. In Chapter 5, we describe our feature selection process, the creation of our prediction model, give a comparison of other common feature selection processes, evaluate our model at different data periodicities, compare our algorithm choice with other common machine learning algorithms, and discuss the results. In Chapter 6, we summarize the scope of our research and identify important conclusions. And finally, in Chapter 7, we identify areas of our research where future work is required.

CHAPTER 2

HARDWARE

Since weather and EMR environmental variables can vary greatly from one location to another, we needed to monitor these variables at the site of the hive for optimal accuracy. To do this, we elected to build our own Weather and EMR Sensing Station that could reliably measure and aggregate data from disparate sensors and persist them so they could be regularly collected for analysis. We also chose this route in an effort to produce a more cost effective means of collecting data that other apiarists could replicate.

We used the "Build your own weather station" tutorial provided by the Raspberry Pi Foundation [34] as a reference while designing, building, and writing some of the weather monitoring portion of the station. As will be discussed later, we improved and built upon this tutorial to build a monitoring station that would meet our needs. This chapter will describe the hardware, source code, station configuration, variables collected, and how the station operates.

2.1 Sensors and Components

Table 2.1 identifies the sensors used by the Weather and EMR Sensing Station, and what was monitored with them.

Sensor/Component	**Variables Monitored**
Bosch BME280	Temperature, Atmospheric Pressure, Relative Humidity
Argent Data Systems Wind / Rain Sensor Assembly	Wind Speed, Wind Direction, Rainfall
Apogee Instruments SP-110-SS Pyranometer	Shortwave Radiation
GQ Electronics EMF-390 Sensor	Radio Frequency, Electric Field, Electromagnetic Field

Table 2.1: Sensors used in the Weather and EMR Sensing Station.

The BME280 sensor is a small, single-board component with sensors to measure the humidity, pressure, and temperature. The operating ranges for each of these sensor measurements are 0 to 100% relative humidity, 300 to 1100 hPa atmospheric pressure, and -40 to +85 °C, with accuracies of ±3 %, ±1 hPa, ±1.0 °C respectively [35]. The RPi.bme280 library for Python 2 or Python 3 can be used for interfacing with the device [36].

Argent Data Systems sells a package containing an anemometer, wind vane, and tipping-bucket rain gauge [37]. The anemometer consists of three equally spaced arms with cups at the end of each. These are attached to a shaft with a magnet at the end that descends into the body of the sensor. As the magnet turns, a reed switch opens and closes due to the magnetic field. At a wind speed of 1.492 MPH, the switch closes once per second [38]. By counting the number of switches per second, the wind speed can be calculated.

The wind vane component consists of a vertical blade attached to a rod with a shaft descending into the body of the sensor. A magnet is attached at the end of the shaft, which is situated at the center of eight reed switches that are all connected to resistors of different values. As the wind vane is turned by the wind, the orientation of the magnet on the shaft causes one or two reed switches to close at a time. An external resistor can be used as a voltage divider, and the output voltage can be measured using an analog to digital converter. Since each switch is connected to a unique resistance, the output voltage will be unique for each of the 16 represented directions each 22.5° apart [38].

The self-emptying tipping-bucket rain gauge consists of a rocker with two cups on each end. The rocker has a hinge connection at the center so only one cup is open to the sky for catching rain at a time. When the cup fills to 0.011" of water, the rocker tips, the rain water dumps out, and the other cup opens to the sky. Each time the bucket tips, a switch is triggered allowing the number of tips to be counted for estimating the rainfall [38].

We purchased the self-powered pyranometer from Apogee Instruments [39]. This silicon-cell sensor measures the 350-1100 nm portion of the solar spectrum of global shortwave radiation, which is approximately 80% of the total range of shortwave radiation [40].

As solar energy enters the sensor, a voltage is produced and fed into an Analog to Digital Converter (ADC) to report the shortwave radiation in Watts per square meter (W/m^2). Each millivolt produced is equivalent to 5 W/m^2, and the sensor can produce a max voltage of about 250 mV to measure up to approximately 1,250 W/m^2 of shortwave radiation [40].

We used the EMF-390 sensor made by GQ Electronics to monitor Electromagnetic Fields (EMF), Electric Fields (EF), and Radio Frequencies (RF) [41]. The EMF portion of the sensor monitors in the X, Y, and Z axes with a range of 0 to ∼500 mG at a resolution of 0.1 per 1 mG. The EF portion of the sensor has a range of 0 to 1000 V/m at a resolution of 1 V/m, and is frequency independent. The RF portion of the sensor has a range of 0.2 µW/m^2 to ∼9999 mW/m^2 at a resolution of 0.01 µW/m^2 and can measure frequencies up to 10 GHz [42].

The EMF-390 sensor was well suited for our needs for several reasons: the range of sensors the EMF-390 provided, hand-held size, and low cost was appealing. This made it easy to house the device in a water-proof enclosure mounted next to the Raspberry Pi computer running the station. The device also allowed for it to be powered via Universal Serial Bus (USB), which the Raspberry Pi could provide. And most importantly, Linux command-line interface software was available, making it possible to log the sensor readings along with the weather observations.

The command-line interface we used for the EMF-390 sensor was written by Davide Dal Farra, and the source code and binaries were hosted on GitLab under the GNU General Public License v3.0 (GPLv3) [43]. This tool could be installed on the Raspberry Pi computer so our software could use it to obtain the sensor readings and log them alongside the weather observations.

Other components used in building the Weather and EMR Sensing Station are shown in Table 2.2.

We used the Raspberry Pi 3 Model B+ as the main controller of the station. This is a single-board computer that can run the Raspbian operating system. It provides 40 General-purpose Input/Output (GPIO) pins that facilitate interfacing with the hardware

Quantity	Part
1	Raspberry Pi 3 Model B+ with Raspbian Buster and Power Supply
1	64GB USB Thumb Drive
1	16-pin DIL/DIP IC Socket
1	MCP3208 12-bit Analog to Digital Converter
2	4.7 kΩ Through-hole Resistor
2	2-pin Male Headers
1	ChronoDot 2.1 (DS3231 Chip) Real Time Clock
1	Adafruit Perma-Proto HAT for Pi Mini Kit - No EEPROM
2	RJ11 Breakout Boards
1	25' 22 AWG Solid Core Insulated Wire
1	3.4 x 3.4 x 2inch (85 x 85 x 50mm) Junction Box
1	5.9 x 4.3 x 2.8inch (150 x 110 x 70mm) Junction Box
1	5.9 x 5.9 x 2.8inch (150 x 150 x 70mm) Junction Box
1	1" x 6" x 24" Mounting Board
1	6' Metal T-Post
1	Vertical T-Post Bracket
1	Horizontal T-Post Bracket
12	1/2" Phillips Screws
1	5' x 3/4" Velcro Roll with Adhesive
1	2"-wide Window Screen Repair Kit Fiberglass Mesh

Table 2.2: Non-sensor components used in building the Weather and EMR Sensing Station.

components. Additionally, it has a built-in 802.11.b/g/n/ac wireless Local Area Network (LAN) [44]. This allowed us to wirelessly connect to the Weather and EMR Sensing Station to regularly collect the data. Finally, the Raspberry Pi comes equipped with four USB 2.0 ports, allowing the EMF-390 sensor and external storage device to be connected to the Pi [44].

The MCP3208 12-bit ADC [45] was used to convert the analog voltage representing the wind direction, and the voltage representing the shortwave radiation from the pyranometer.

Since the Raspberry Pi doesn't have a built-in Real Time Clock, if it's disconnected from the internet, it can't accurately keep track of the time. For this reason, we incorporated the ChronoDot 2.1 Real Time Clock into our system. This device is temperature-compensated and has a drift of approximately 1 minute per year [46]. This allowed the Raspberry Pi to refer to the time being kept by this device instead of the internet so our log records could be accurate.

To keep the hardware organized and the wiring less prone to user error, we chose to mount the various hardware components and wire them together using the Adafruit Perma-Proto Hardware Attached on Top (HAT) board. This board consisted a grid of holes and pads through which we could solder the hardware components. The board also has a header that can slide onto the Raspberry Pi's GPIO pins which were then routed to specific holes on the board. This allowed us to be able to easily swap out Raspberry Pi devices, if needed, without having to redo any of the wiring. It also made the device much more compact.

Most of the remaining components in Table 2.2 were used to house and deploy the Weather and EMR Sensing Station to the bee hive site. To keep the EMF-390 sensor, Raspberry Pi, and BME-280 sensors from being exposed to the elements, we put them in waterproof junction boxes. We chose to put the BME-280 sensor in its own junction box so the values wouldn't be skewed by the heat generated by the Raspberry Pi. We also put a hole at the bottom of the box and covered it with a fiberglass screen so the air could more freely flow around the sensor for more accurate readings. The screen helped prevent wasps and other insects from interfering. Additionally, we chose to put the EMF-390 sensor in its own box with the intent that its readings wouldn't be as skewed by any EMR generated by the Raspberry Pi.

We then screwed these junction boxes to a 1" x 6" x 2' pine board, and used a vertical T-post bracket to hang the board on a metal T-post. When mounting the junction boxes, it was important to have the junction boxes mounted on the North face of the wooden mounting board. This made it so the sensors could obtain more accurate readings because they would mostly or entirely be in the shade throughout the day due to the northern latitude of the bee hive site in Logan, Utah.

We then used a horizontal T-post bracket hung on a vertical metallic T-post and mounted the pyranometer to it. This was followed by the anemometer, wind vane, and rain gauge assembly vertically attached to the top of the T-post. In doing this, it was important that the rain gauge be pointed North, the pyranometer pointed South, and the wind sensors pointed West and East so as to prevent any shadows produced by the station from crossing

the pyranometer. Figure 2.1 shows the assembled Weather and EMR Sensing Station.

(a) (b)

Fig. 2.1: The North facing side of an assembled Weather and EMR Sensing Station at a test site with the junction boxes open (2.1a) and the South facing side (2.1b) of an assembled Weather and EMR Sensing Station at the data collection site.

2.2 Software

As mentioned previously, we were able to reference the "Build your own weather station" tutorial for some of the software components of the weather station. This provided us with a starting point for the weather station, which we were able to customize and add to so we could have a station that would meet our needs. The Weather and EMR Sensing Station code used for this paper has been made available to the public, and can be found on GitHub at https://github.com/lightningWhite/WeatherAndEMFSensingStation. Additional documentation regarding the station can be found in the readme.md file in that repository. The code is written in Python 3, and some bash scripting was used as well.

At a high level, the code in the above-mentioned repository can be cloned to a Raspberry Pi running the Buster version of Raspbian. A few manual steps (documented in the readme.md file) need to be performed, after which the install.sh script can be executed to

perform the bulk of the installation. Once the Raspberry Pi is restarted, the Weather and EMR Sensing Station begins monitoring and logging the weather and EMR automatically.

The weather station has a main loop that creates a Comma Separated Values (CSV) file named by the date and time the station began monitoring. This file will either be created on an attached USB storage device, or, if one isn't present, on the Secure Digital (SD) Card where the operating system is installed. This file is where the Weather and EMR values are collected. An additional file is also be created where logs of the station's operation are stored.

After the data and log files are created, the station's main program loop begins. Since some of the variables monitored change rapidly, this loop consists of a nested loop that accumulates readings for the wind speed, wind direction, and the RF, EF, and EMF readings. This allows these rapidly changing values to be averaged over the duration of the log interval of 15 minutes for more accurate measurements. It also allows for maximums to be calculated and logged for those variables. The values of all other variables (temperature, atmospheric pressure, relative humidity, rainfall, and shortwave radiation) are obtained at the end of each 15-minute accumulation period.

At the end of every 15 minutes, the averaged values, calculated maximums, and other readings are appended to the main data CSV file. Before the values are written, the primary data file gets backed up to reduce the possibility of data corruption during a power outage. This loop continues, and at midnight the accumulated rainfall gets reset to zero. Also, at the end of each month, the log file gets cleared.

Although we were able to refer to the Raspberry Pi Foundation's tutorial for some of the sensor interface code, we ended up having to rewrite and customize much of it to suit our project requirements. Some of these changes included fixing various bugs, modifying the overall program structure, adding additional sensor interface code, improving data and system logging capability, increasing sensor resolution, making the station more robust to power outages, allowing the station to operate disconnected from the internet, and improving the installation and setup process.

We were able to use the RPi.bme280 Python library [36] to interface with the BME-280 temperature, pressure, and humidity sensor. Since the atmospheric pressure sensor is factory-calibrated for readings at sea-level, the readings needed to be calibrated for different elevations. To do this, we added a calibration variable that would need to be set appropriately when deployed. The calibration value can be obtained by running the station to obtain a pressure reading, looking up the pressure reading of a reliable station at the same elevation, and calculating the difference between the two. Once the calibration value gets set with this calculated value (either positive or negative), it gets added to the default sensor reading before it is logged.

The wind speed sensor code is fairly simple. The gpiozero Python library [47] provides a Button module that can be used to monitor the GPIO pin connected to the wind speed sensor wire. An interrupt is configured (using the Button module) so whenever the switch is triggered inside the wind speed sensor, a function is called which increments the value of a variable. The counts are accumulated over a period of time and then multiplied by the diameter of the rotating cups to calculate the distance traveled by a cup during that time. The wind speed is then calculated in Miles per Hour (MPH).

The rain gauge code was very similar to that of the wind speed sensor. The Python Button module was used to monitor the GPIO pin connected to the switch in the rain gauge sensor. Whenever the switch is triggered, the amount of water held by the bucket in inches (0.011) is added to an accumulation variable, which gets cleared every day at midnight.

To obtain the direction the wind is coming from using the wind direction sensor, the MCP3208 module of the gpiozero Python library was used. This module provided a simple means of utilizing the MCP3208 ADC to convert the different analog voltages into digital values that the Raspberry Pi could use. When a voltage was received by the ADC via the MCP3208 Python module, it was matched to the nearest voltage that mapped to one of the 16 valid directions stored in a Python dictionary. The associated wind direction was then returned.

The pyranometer interface code for detecting shortwave radiation from the sun also

made use of the MCP3208 Python module. The voltages produced by the sensor that represent different levels of Shortwave Radiation were fed into the ADC to convert the analog signal to a digital value. The Raspberry Pi could then multiply the digital value obtained by the sensor's calibration factor of 5.0 for every millivolt detected. The resulting value represented the Shortwave Radiation value.

In order to interface with the EMF-390 sensor, a connection first had to be established. Since the EMF-390 sensor and an external USB storage device could be connected at the same time, code had to be written to determine which port the sensor was connected to. To do this, a function was written that would obtain all connected USB devices. Then it would loop through each device and attempt to query the EMR readings from each device until it succeeded. If the readings came back successfully for a device being tested, it would be used for the remaining lifetime of the program.

Once the correct device was determined, the Weather and EMR Sensing Station code would invoke the emf390cli [43] application to obtain the command line output containing the EMR readings. This included readings for the RF and associated frequencies in various units, EF, and EMF. We then parsed the CSV text output, obtained the desired values, and converted the frequency units to megahertz (MHz) for uniformity. These values were then returned to the main program loop where they could be accumulated over each 15 minute interval, averaged, and logged.

The final piece of code written for the Weather and EMR Sensing Station is the installation and startup scripts. These were written to make it easier and more repeatable to install and set up the station.

In general, the installation script copies an init.d script into the /etc/init.d directory that gets called when the Pi starts. This script invokes a tmux session, calls another script to source and activate the Python virtual environment to make all of the necessary dependencies available, starts the main program, and verifies that it started successfully. By starting the station in a tmux instance, users can later connect to and disconnect from the terminal to see the real-time output of the station without killing the application. This

comes in very handy when doing routine station checks after deployment. Also, having the station start up automatically is critical for robustness against power outages for minimal data loss.

The installation script also creates a mount point for the USB external storage device and adds it to the fstab so it will be mounted automatically when the station starts. Additionally, it configures a default network to which the Pi will automatically attempt to connect. This becomes very useful for data collection or station checks. For example, by using a smart phone to start a WiFi hotspot with the correct network name and password, the Pi will automatically connect to it making it possible to transfer the data to the phone or even somewhere online. By storing the data in a CSV format, it's particularly easy to collect the data, as opposed to a database for example. It also makes it possible to communicate with the device via a Secure Shell (SSH) connection (which is also enabled by the installation script) to the device for routine maintenance or updates.

Other important configurations performed by the installation script include creating the required data and log directories, enabling and starting SSH, and configuring the Pi to synchronize with the Real Time Clock (RTC). Setting the time of the RTC initially is a manual step, but after running the installation script, the Pi will synchronize with the RTC. This is especially important when the Pi is running disconnected from the internet. Without this, the Pi's clock will drift over time, and if a power outage occurs, the time will be incorrect when it starts up again.

As mentioned before, all of the source code is available on GitHub. The readme documents how to install it, how it operates, additional information on how to interface with it to collect the data and maintain it, and helpful details about each of the files used. Further, all of the hardware wire connections are documented, along with photos and tips for constructing the station.

2.3 Collected Variables

The Weather and EMR Sensing Station logs 29 different variables to a CSV file. These include direct readings, averages, and maximums. Table 2.3 lists the column name and a

description for each variable collected. All average and maximum values recorded pertain to the previous 15 minute interval, and readings to calculate those values were sampled every 10 seconds during that time. The Precipitation value represents a continuous accumulation of rain fall since midnight, where the latest value is recorded every 15 minutes with the other variables. All other values are determined and recorded at the end of the 15 minute interval.

While the EMF-390 sensor was running for our experiments, it was configured to monitor the frequencies between 240 MHz and about 1040 MHz. The variables pertaining to RF Watts and RF Density in Table 2.3 are in relation to frequencies within that range. However, the Avg. Total Density and Max Total Density variables pertain to the signals from all sources with frequency bands between 0.01 GHz to 10 GHz.

Column Name	**Description**
Record Number	Integer starting from one, incremented for each new row
Time	Date and time the row was recorded using YYYY-mm-dd HH:MM:SS format
Temperature (F)	Ambient temperature in degrees Fahrenheit
Pressure (mbar)	Atmospheric pressure in millibars
Humidity (%)	Relative humidity in percent
Wind Direction (Degrees)	Average direction the wind is coming from in degrees from North (0-359)
Wind Direction (String)	Average direction the wind is coming from as a string (e.g. W)
Wind Speed (MPH)	Average speed of the air moving past the sensor in Miles per Hour
Wind Gust (MPH)	Maximum wind speed in Miles per Hour
Precipitation (Inches)	Accumulated precipitation since midnight in inches

Shortwave Radiation (W mˆ(-2))	Global shortwave radiation from the sun (direct beam and diffuse) incident on a horizontal surface [40] measured in Watts per square meter
Avg. RF Watts (W)	Average power in Watts of the RF signals detected
Avg. RF Watts Frequency (MHz)	Average frequency in megahertz of the signals associated with the average RF power level detected
Peak RF Watts (W)	Maximum RF power level detected in Watts
Frequency of RF Watts Peak (MHz)	Frequency in megahertz associated with the maximum RF power level detected
Peak RF Watts Frequency (MHz)	Maximum frequency in megahertz detected while monitoring the RF power
Watts of RF Watts Frequency Peak (W)	Power in Watts associated with the maximum RF frequency detected
Avg. RF Density (W mˆ(-2))	Average RF power flow per unit area in Watts per square meter [48]
Avg. RF Density Frequency (MHz)	Average frequency in megahertz of the signals associated with the average RF density detected in megahertz
Peak RF Density (W mˆ(-2))	Maximum RF density detected in Watts per square meter
Frequency of RF Density Peak (MHz)	Frequency in megahertz associated with the maximum RF density level detected
Peak RF Density Frequency (MHz)	Maximum frequency in megahertz detected while monitoring the RF density

Density of RF Density Frequency Peak (W mˆ(-2))	Density in Watts per square meter associated with the maximum RF frequency detected
Avg. Total Density (W mˆ(-2))	Average total RF power flow per unit area in Watts per square meter [48] from all sources with frequency bands between 0.01 GHz to 10 GHz
Max Total Density (W mˆ(-2))	Maximum total RF power flow per unit area in Watts per square meter [48] from all sources with frequency bands between 0.01 GHz to 10 GHz
Avg. EF (V/m)	Average strength of detected electric field in volts per meter
Max EF (V/m)	Maximum strength of detected electric field in volts per meter
Avg. EMF (mG)	Average strength of detected electromagnetic field in milligauss
Max EMF (mG)	Maximum strength of detected electromagnetic field in milligauss

Table 2.3: Column names and descriptions of the variables collected and stored in a CSV file by the Weather and EMR Sensing Station.

Since the weather variables are commonly used, we will not go into further detail about them here. However, later sections will delve deeper into how they are interrelated while exploring their relation to honey bee activity.

Electromagnetic radiation can be defined as "waves of electric and magnetic energy moving together (i.e., radiating) through space at the speed of light" [48]. While the electromagnetic spectrum spans a wide range of frequencies, the sensor used in our research was capable of monitoring RF signals between 0.01 GHz up to about 10 GHz [42]. It was

also capable of individually monitoring both the electric component (EF) and the magnetic components (EMF) of detected fields.

Our research utilized three components of the RF signals detected: Frequency, power, and density. The frequency of a signal represents how often during a given period of time the wave oscillates, and the power component represents the strength of the RF signal. The density represents the RF power flow per unit area at a given location [48].

An electric field, represented by the EF variable in our data set, is created when an object has either a positive or negative charge. This charge object can exert a force on other charged objects. The EMF-390 sensor measures the force created by a charged body in volts per meter, which is equivalent to Newtons per Coulomb.

An electromagnetic field is a field that consists of an electric component, as described above, and a magnetic component. The magnetic component of the field is created by flowing electrons.

The ambient EMR signals as detected by the EMF-390 sensor at our hive location in Logan, Utah, correspond to many different sources. According to the U.S. Department of Commerce as of 2016, some uses of the frequencies that would fall in the 240 MHz to 1040 MHz range (potentially captured by the RF Watts and RF Density variables in Table 2.3) include things such as cell service (mobile), aeronautical communication, satellite communication and earth exploration, meteorological monitoring, amateur radio, radio astronomy, space research, and television broadcasting [49]. The larger 0.01 GHz to 10 GHz range, as potentially captured by the Avg. Total Density and Max Total Density variables in the aforementioned table, include extended usage of all those previously mentioned, as well as others such as FM radio [49]. It also includes things like Bluetooth, WiFi, Microwave Ovens, and smart meters [42].

CHAPTER 3

DATA COLLECTION AND CURATION

3.1 Data Collection

We deployed two Weather and EMR Sensing Stations at a six-hive apiary in Logan, Utah, USA. The location is near 41.73698° latitude, -111.833836° longitude at an elevation of approximately 4,457 feet above sea level. Since the hive site was on private property, the exact location is not provided.

The hive site where data collection occurred consisted of a wooded backyard within a residential neighborhood as shown in Figure 3.1. The two Weather and EMR Sensing Stations were deployed in a small clearing not more than 60 feet away from the furthest of the six hives. All of the hives resided within a 40 feet by 60 feet area.

We used BeePi monitors developed by Kulyukin et al. [20] to individually monitor each of the bee hives. The BeePi monitor consists of a camera, microphone, and temperature sensor, each connected to a Raspberry Pi residing in a Langstroth super box placed on top of the hive. The eight-megapixel camera captured 30 second video clips from the top super of the hive every 15 minutes throughout the day.

Both Weather and EMR Sensing Stations and all six BeePi monitors recorded observations every 15 minutes. While the Weather and EMR Sensing Stations operated continuously, the BeePi monitors ceased capturing video data each night from midnight until 7:00 AM. This was due to decreased visibility at night as well as the absence of honey bee foraging activity during those hours.

The Weather and EMR Sensing Stations were deployed on May 16, 2020. At the time, five hives were being monitored by BeePi devices. One additional hive was added to the site, and data collection for that hive began on May 22, 2020. Data collection continued until October 26, 2020.

Fig. 3.1: Weather and EMR Sensing Stations, BeePi Monitors, and bee hives at the data collection site in Logan, Utah.

During the season, we experienced some hardware failures of both the BeePi monitors and the Weather and EMR Sensing Stations. Since at least one of the two Weather and EMR Sensing Stations was running at any given time throughout the season, we hoped to combine the records from each station to produce a single continuous data file. However, we later discovered that the EMF-390 sensors reported values that weren't calibrated the same with one another. Thus, we chose to keep the station data separate and handle any data gaps that occurred.

Table 3.1 and Table 3.2 shows the specific data ranges of the data used from the BeePi Monitors and Weather and EMR Sensing Stations respectively, along with any periods where data is missing.

It should be noted that on July 25, 2020, data collection for the hive monitored by

Hive Identifier	Data Range	Missing Data Ranges
R_4_5	5/16/2020 - 10/26/2020	N/A
R_4_7	5/16/2020 - 10/26/2020	7/18/2020 - 8/24/2020
R_4_8	5/16/2020 - 7/25/2020	N/A
R_4_10	5/16/2020 - 10/26/2020	7/18/2020 - 8/24/2020
R_4_11	5/16/2020 - 10/26/2020	N/A
R_4_14	5/22/2020 - 10/26/2020	N/A

Table 3.1: Date ranges for the data collected by each hive's BeePi monitor, as well as periods where data is missing due to hardware failures. Hive identifiers correspond to the IP address of the BeePi monitor for each hive.

BeePi monitor R_4_8 ended. This was because the hive became too tall for the BeePi monitor to function properly due to the addition of another Langstroth super to satisfy the needs of the growing colony.

Station Identifier	Data Range	Missing Data Ranges
1	5/16/2020 - 10/25/2020	2020-06-17 00:37 - 2020-06-20 17:43 2020-07-20 17:46 - 2020-07-20 18:05 2020-07-22 22:27 - 2020-07-23 18:19 2020-07-31 11:05 - 2020-07-31 11:26 2020-08-08 10:20 - 2020-08-08 11:06 2020-08-12 11:10 - 2020-08-15 19:35 2020-10-07 01:48 - 2020-10-10 11:19 2020-10-23 04:13 - 2020-10-24 17:36
2	5/19/2020 - 10/26/2020	2020-05-20 23:43 - 2020-05-23 17:11 2020-06-20 17:38 - 2020-06-20 18:03 2020-07-31 10:57 - 2020-07-31 11:26 2020-08-21 10:32 - 2020-09-08 17:10 2020-10-18 20:29 - 2020-10-24 18:03

Table 3.2: Date ranges for the data collected by each Weather and EMR Sensing Station, as well as periods where data is missing due to hardware failures. A gap between records larger than about 15 minutes is considered missing data.

The Weather and EMR Sensing Stations and the BeePi monitors were all equipped with external storage devices. This data was regularly collected by either swapping out storage devices or by remotely transferring the data. It could then be taken back to the lab for further processing and experimentation.

3.2 Data Preprocessing

The data collected by the Weather and EMR Sensing Stations and the BeePi monitors needed to be preprocessed and combined before it could be used in our prediction models.

After producing several CSV files from each Weather and EMR Sensing Station, we combined them into a single master CSV file for each station. It should be noted that certain periods of the collected raw data had to be corrected before it was added to the master data file. One correction made was due to a six minute power outage. When the station came back up, a bug (now resolved) prevented the Pi from synchronizing with the RTC. Instead of using the correct time, the Pi reported a time 54 minutes behind what it should have been. To correct the timestamps, we added 54 minutes to the timestamps of the records from when the power outage occurred to when the problem was discovered and corrected. The corrected data spans about one week from 7/31/2020 until 8/08/2020.

We also had to correct the atmospheric pressure readings on the raw data spanning 8/21/2020 until 10/07/2020. A software update was installed on one of the stations that contained an incorrect calibration value for the barometer. This resulted in pressure readings 6.226 millibars higher than they should have been. To correct the data, we subtracted 6.226 from each of the atmospheric pressure readings during the above mentioned time period.

Once we had a single file for each station containing the continuous data, we began preprocessing it. First, we removed any extra spaces after the commas in the CSV file to make it more uniform. Then we rounded the time stamps either up or down to the nearest quarter hour. The rounding of the timestamps needed to be performed since the Weather and EMR Sensing station recorded the observations every 15 minutes from when the station was started. This made it so the BeePi monitor data that is captured on each quarter-hour could be merged effectively with the newly aligned timestamps.

We also detected significant outliers in the EMR data that needed to be handled. To remedy these, we replaced all values that fell outside the third standard deviation of the data in each individual EMR column with an interpolated value. We didn't detect any values

in the Weather data that could be considered outliers, so we didn't apply this function to those data.

Since we collected video clips every 15 minutes using the BeePi monitors on each hive, we needed to process those video clips to produce directional bee motion counts that could be paired with the weather and EMR data. To do this, we used an algorithm based on digital particle image velocimetry proposed by Kulyukin and Mukherjee [21] [50], and processed a total of 61,739 videos over the course of several months.

To briefly describe how this proposed algorithm operates, it first takes two consecutive video frames and compares them. A correlation formula is then used to find out how a point in the first image corresponds to a point in the second image. Then a displacement vector can be calculated between the points with highest correlation, which represents how the particles may have moved from one frame to the next. Displacement vectors are calculated for each moving particle in the region to create vector fields that can be used to estimate possible bee motion patterns [21] [50].

Since several motion vectors may correspond to a single bee's motion, they need to be condensed into single vectors to produce accurate motion counts. To do this, correlation matrices are generated for all motion vectors within a configured interrogation window size. Signal to Noise Ratios (SNR) are then calculated between the first and second highest correlations peaks, and vectors with associated SNR values that don't meet a certain threshold are deemed spurious and removed. The remaining vectors pertaining to a bee are replaced by a weighted average of the neighboring vectors. By representing each bee's motion by a single vector, they can be labeled as one of three directions (incoming, outgoing, or lateral), and the sums of each can be calculated [21] [50].

We used the algorithm proposed by Kulyukin and Mukherjee [21] [50] to operate on a directory of videos captured by the BeePi monitor. The script produced an individual CSV file for each 30 second video, and was named by the date and time the video was captured. Each CSV file contained a row specifying the upward, downward, lateral, and total bee motion count values as estimated by the algorithm for each frame captured during the 30

second video.

After producing a series of CSV files, one for each video, we consolidated the data from those files into a single CSV file for each hive. A hive's master CSV file contained columns for the time stamp, upward motion, downward motion, lateral motion, and total motion. Each row of the new CSV file contained the sums of each video's directional columns. The time stamps were pulled from the filename of the CSV file being processed, and the seconds of the time stamp were rounded to zero since they were all one second after the quarter hour. This format enabled the directional bee motion count data to be paired up with the weather and EMR data.

Once the weather and EMR data files and bee motion count files were prepared as described, they could then be merged. To do this, our script first determines the latest start time and earliest end time between the two files to be merged. Then a new base CSV file is created with a record number column and a time stamp column, and rows consisting of timestamps incremented by 15 minutes ranging from the latest start time to the earliest end time.

After creating the base file, the weather and EMR data are added for each matching time stamp in the base file. Any potentially missing rows are filled with Not a Number (NaN) values for each data column. Then the directional bee motion count data can be added for each matching time stamp as well. Since the directional bee motion count data only covered the hours from 7:00 AM until midnight, the missing hours were filled with NaN values. This merging process was performed 12 times to produce a merged weather, EMR, and bee motion count CSV file for each of the six individual hives and the two Weather and EMR Sensing Stations.

After analyzing the data, we realized that the Shortwave Radiation data collected by our station was skewed due to the shadows produced by the trees at the bee hive yard. Since this variable has been shown to be a critical predictor variable in other research [24] [23], we felt it essential that we have comparative shortwave radiation data in our experiments.

To remedy this, we decided to use shortwave radiation data collected by a nearby Utah

Climate Center weather station located on USU campus. This station was approximately two miles away from the hive site, and about 300 feet higher in elevation. The USU weather station recorded observations once every hour. We were able to download the data from the Utah Climate Center's website for the station [51].

Since the data from the Utah Climate Center was hourly data, we expanded it into 15 minute increments by assigning each 15 minute increment within the hour to the hourly value. We also rounded the hour value so the minutes would be on the hour exactly instead of at 59 minutes. We then added this shortwave radiation data as an additional column to the master CSV file for each hive.

After these steps were performed, the final preprocessed CSV file included the columns documented in Table 3.3 in addition to those previously documented in Table 2.3.

Column Name	**Description**
BeeMonitorID	The last two sections of the IP address of the BeePi monitor
UpwardMotion	Number bee motions in the range [11°, 170°] for a 30 second time period [50]
DownwardMotion	Number bee motions in the range [-11°, -170°] for a 30 second time period [50]
LateralMotion	Number bee motions in the ranges [-10°, 10°], [171°, 180°], [-171°, -180°] for a 30 second time period [50]
TotalMotion	Sum of the UpwardMotion, DownwardMotion, and LateralMotion values for a 30 second time period
Shortwave Radiation USU (MJ/mˆ2)	Shortwave solar radiation [52] in megajoules per square meter

Table 3.3: Directional bee motion count columns and Utah Climate Center USU weather station column.

The software written as part of this research for preprocessing the data has been made available to the public, and can be found on GitHub at `https://github.com/lightningWhite/BeeMotionWeatherEMFDataHandling`, or readers are encouraged to contact the authors for access to the software.

3.3 Curated Data Sets

As a result of collecting and preprocessing the data from the two Weather and EMR Sensing Stations and each of the six BeePi monitors, we have 12 curated data sets that are made available to the public. This has been done so interested parties can replicate our findings, and so accuracy benchmarks can be established. Table 3.4 shows the name and brief details of each file. These data sets can be downloaded from https://github.com/lightningWhite/WeatherEMRAndBeeMotionCountData-2020, or readers are encouraged to contact the authors for the availability of the data.

Filename	Rows	Date Range	Size
R_4_5_s1_2020.csv	15,509	05/16 - 10/25	3.4 MB
R_4_7_s1_2020.csv	15,509	05/16 - 10/25	3.4 MB
R_4_8_s1_2020.csv	6,721	05/16 - 07/25	1.5 MB
R_4_10_s1_2020.csv	15,509	05/16 - 10/25	3.4 MB
R_4_11_s1_2020.csv	15,509	05/16 - 10/25	3.4 MB
R_4_14_s1_2020.csv	14,970	05/22 - 10/25	3.3 MB
R_4_5_s2_2020.csv	15,299	05/19 - 10/26	3.2 MB
R_4_7_s2_2020.csv	15,299	05/19 - 10/26	3.1 MB
R_4_8_s2_2020.csv	6,418	05/19 - 07/25	1.5 MB
R_4_10_s2_2020.csv	15,299	05/19 - 10/26	3.1 MB
R_4_11_s2_2020.csv	15,299	05/19 - 10/26	3.2 MB
R_4_14_s2_2020.csv	15,063	05/22 - 10/26	3.1 MB

Table 3.4: The file names, the total number of rows, the start and end dates in the year 2020, and the file sizes in megabytes (MB) for each data set file. Note that these files contain the data gaps documented in Tables 3.1 and 3.2.

CHAPTER 4

DATA ANALYSIS

After collecting and preprocessing the data, we proceeded to analyze the data at a surface level to better understand how the weather and EMR variables correspond to honey bee total motion. We also sought to identify relationships between variables to know how they could best be used to predict bee motion while avoiding skewed results. The findings in this chapter heavily influenced the experiments we performed later in our research, as well as our results and conclusions.

4.1 Data Correlation

To begin our analysis, we used Pearson's correlation coefficient to compare the covariance of each column with every other column. We did this by generating a correlation matrix represented as a heat map, which can be found in Figure 1. Here we show a simplified, annotated heat map for the total motion column alone in Figure 4.1. This analysis allowed us to identify positive and negative correlations of interest that could be used in predicting bee total motion. Additionally, analyzing the correlations is an effective way to quickly understand how the data is interrelated as well as gain a premonition as to what variables may be most useful in predicting bee activity. This information could also be valuable to entomologists in understanding how environmental conditions correspond to bee activity. All of the values, figures, and insights provided in this chapter come from the weather and EMR data collected by Station 1 paired with the R_4_5 hive's bee motion data.

We will provide some observations and insights regarding the most noteworthy correlations between the variables collected by the Weather and EMR Sensing Station and the bee total motion data collected by the BeePi monitor. These correlations can be seen in Figure 4.1. Insights will also be given on various correlations among the weather and EMR variables themselves. While select plots have been included in this chapter, others can be

found in the Appendix.

4.1.1 Weather and EMR Variable Correlations with Bee Total Motion

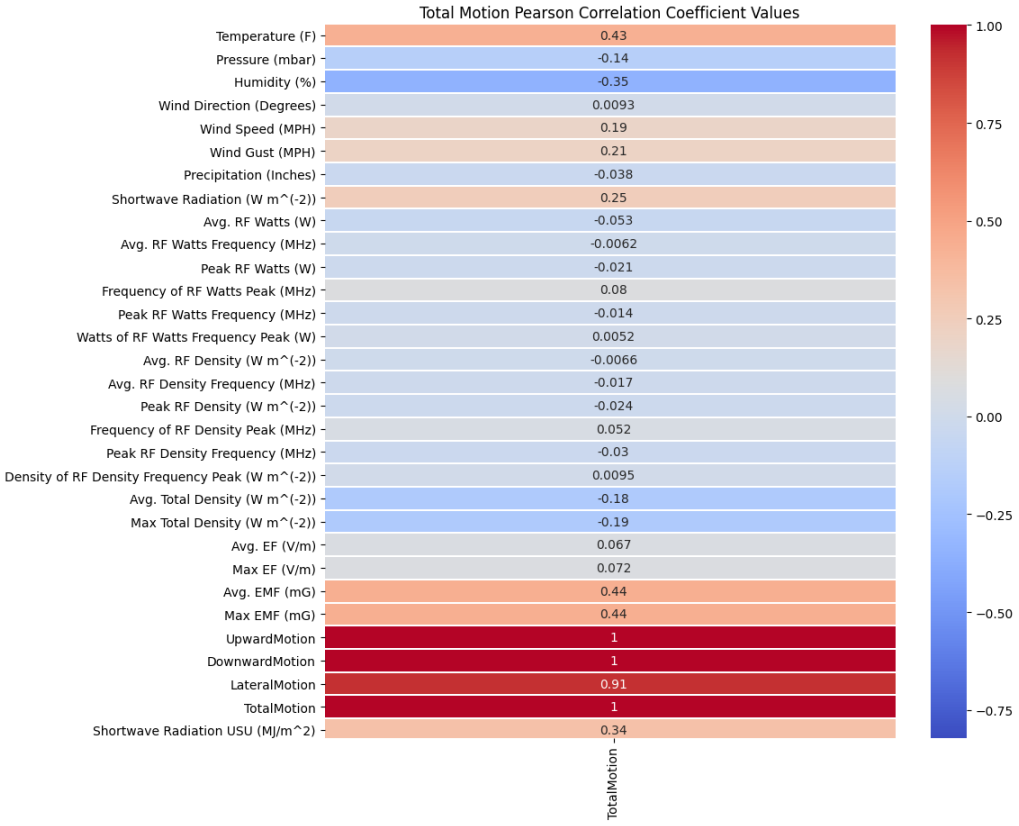

Fig. 4.1: The Pearson correlation coefficient values of the Total Motion column with every other column.

Temperature and Total Motion

The temperature has a strong positive correlation with bee total motion of 0.43. As the temperature rises and falls each day, the bee motion also typically rises and falls in conjunction with it. Figure 2 shows that the peak bee activity typically occurs either shortly before or after the peak daily temperature. These observations concur with findings of previous research [24] [25] [16] [23] [26] [27] [29].

Atmospheric Pressure and Total Motion

The atmospheric pressure has a negative correlation with bee total motion of -0.14. Inspection of the atmospheric pressure plot in Figure 3 indicates that it follows a diurnal pattern where it is typically the highest in the late morning and lowest in the late evening. These daily fluctuations of the pressure are usually due to the daily warming of the atmosphere by the sun [53]. It should be noted that the bee motion often peaks while the pressure is falling. Thus, this characteristic could potentially be utilized in predicting bee motion.

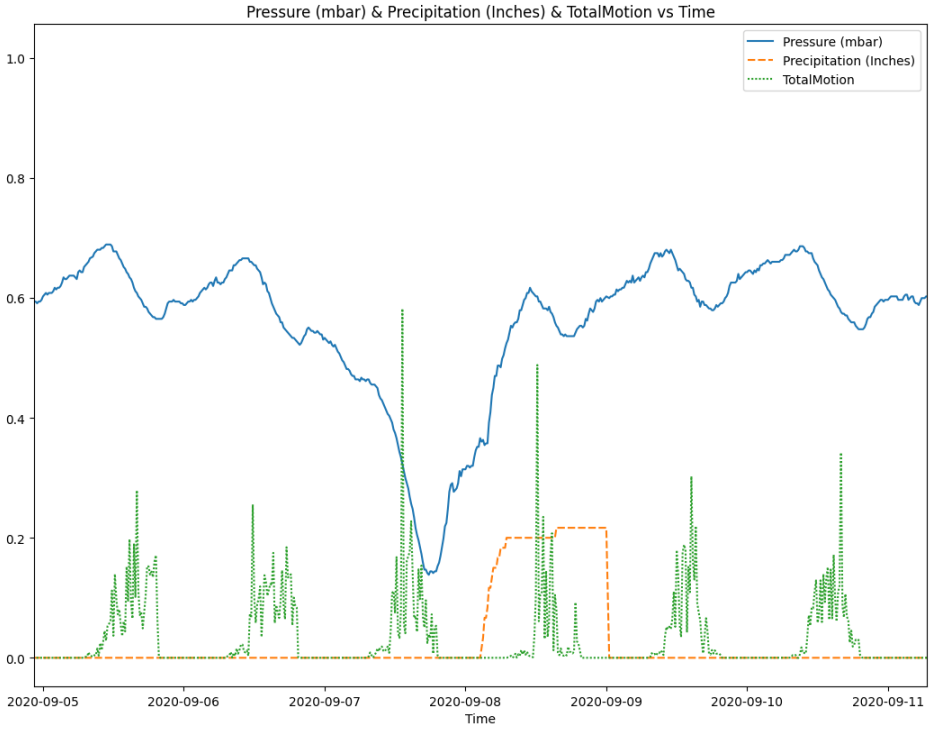

Fig. 4.2: Normalized Atmospheric Pressure, Precipitation, and Total Motion plotted against time. Note that bee total motion spikes to higher-than-normal values during the steep atmospheric pressure drop. Precipitation was recorded during the next day, indicating a storm.

It is also known that atmospheric pressure usually decreases before a storm arrives.

These changes in pressure are usually larger than the normal diurnal drops. In some of these instances, as shown in Figure 4.2, it's observed that bee motion may spike during these deep, sharp pressure drops before a storm arrives. This plot shows that the barometric pressure falls deeper and at a greater rate than the cycles observed on a normal daily basis. It's also observed that rain accumulation was detected shortly after the pressure drop, indicating that a storm followed the pressure change.

However, precipitation doesn't occur very often during the summer in Logan, Utah, so additional data would be needed to explore this further to determine whether bees respond to steep drops in atmospheric pressure to, in effect, "predict" the weather. Regardless, machine learning algorithms could potentially key off of this distinct pattern to aid in predicting bee motion.

Humidity and Total Motion

Relative Humidity exhibits a strong negative Pearson correlation with total motion of -0.35. It should be noted that there is a strong inverse correlation between temperature and relative humidity as observed in Figure 1. Relative humidity can be defined as "the amount of atmospheric moisture present relative to the amount that would be present if the air were saturated" [54], and it "is a function of both moisture content and temperature" [54]. Since warm air can hold more moisture, this means that if the moisture content stays the same while the temperature decreases, the relative humidity would increase.

This interdependence makes it difficult to identify whether the relative humidity or the temperature exhibits greater influence on the bee activity. It may have been beneficial to include a sensor that could measure absolute humidity to better see the effects of air moisture on bee activity. However, the relative humidity variable can still give insight into whether the bees are responding to the moisture saturation of the air.

Figure 4.3 shows how the relative humidity varies inversely with the temperature and bee total motion. When the relative humidity decreases, the bee motion and temperature readings increase.

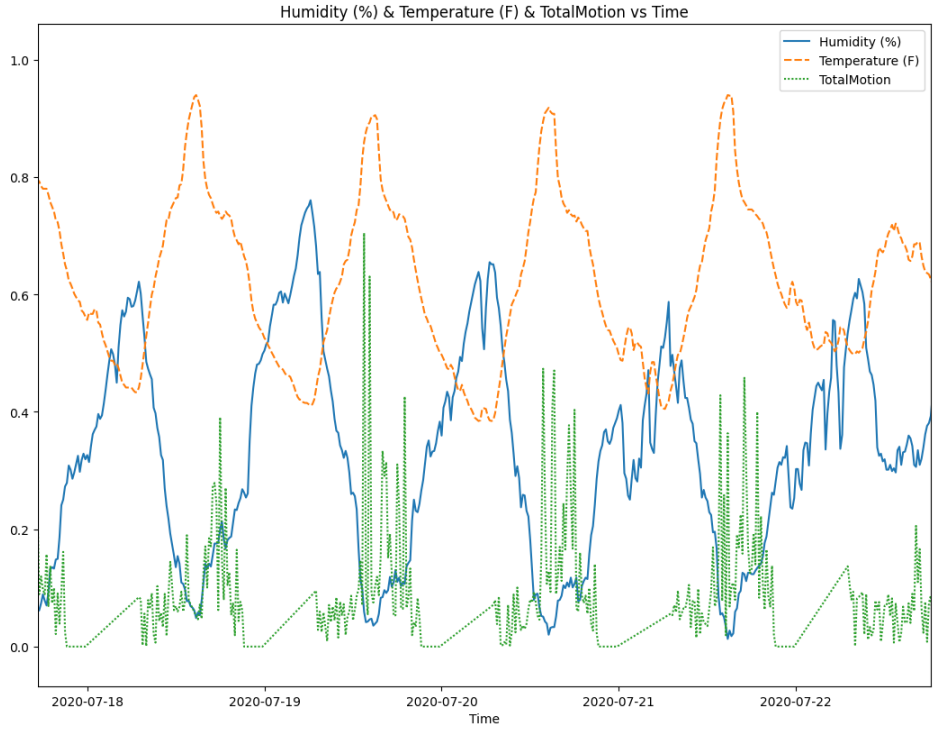

Fig. 4.3: Normalized Relative Humidity, Temperature, and Total Motion plotted against time. Note how the relative humidity decreases while the temperature and bee total motion increase. Fluctuations in the air's moisture content may provide valuable predictive information.

Wind Speed and Total Motion

The average wind speed was calculated to have a positive correlation with bee total motion of 0.19. Figure 4 shows that the wind speed is generally higher during the day, and lower during the night. Since bee foraging activity ceases at night, this correlation would naturally arise.

Nevertheless, a closer inspection of the plots reveal that there may be a more localized negative correlation of bee activity to wind speed. Figure 5 shows that, on occasion, when the wind speed increases, bee total motion decreases or even ceases. It may be possible that there is some wind speed threshold at which it begins to negatively impact bee motion. This finding aligns with that of Hennessy et al. [30] where they observed that if the wind

speed was moderate, foraging activity would stop. This relationship with bee activity may be able to be utilized by a bee motion prediction model to account for windy days.

Precipitation and Total Motion

The Precipitation variable exhibits a very slight negative correlation of -0.038 with bee Total Motion. This overall correlation is likely minor do to the limited precipitation that was received during the season.

Figure 6 shows that bee motion drops significantly when rain occurred at times bee activity would normally be high. This is indicated in the plot by the abrupt cessation of bee motion at the same time that the precipitation value increased. Later in the day when the precipitation plot line leveled out (indicating that the rain stopped), bee motion resumed. Note that the precipitation value is an accumulation metric that clears at midnight each day.

We also observed that on occasion it appeared that the bees exhibited higher levels of activity after it rained. This can be seen in Figure 4.2 by the higher-than-normal spike in activity the day following a rain storm. This increased activity after rain was also observed by Devillers et al. [26].

If more rain was received during the season, this negative correlation would likely have been shown to be stronger. However, this relationship will likely be key in being able to predict bee motion on rainy days.

Shortwave Radiation and Total Motion

The Shortwave Radiation USU variable shows a positive correlation with bee total motion of 0.34. This correlation score is higher than the 0.25 value reported for the Shortwave Radiation data collected by our Weather and EMR Sensing station. While the shortwave radiation increased, we observed that bee motion also increased.

Upon inspecting the plots of the Shortwave Radiation collected by our station, we observed that the data had many more peaks and troughs throughout the day compared to the data collected by the Utah Climate Center's station. This led us to believe that these

discrepancies in our station's data were caused by shadows passing over our sensor that were created by the trees at the data collection site. Figure 7 shows a comparison of the data collected by the two sensors. This ultimately led us to use the Shortwave Radiation data collected by one of the Utah Climate Center's stations in our experiments rather that our station's data.

The positive correlation we observed between the shortwave radiation and bee activity agrees with previous research [24] [23]. As can be seen in Figure 8, a drop in Shortwave Radiation in the afternoon of June 3rd coincides with a drop in bee total motion. We also observed that this drop occurred at about the usual time of peak bee activity, as shown by the bee motion the day before. This relationship will likely prove to be a very useful predictor of bee motion.

Avg. Total Density and Total Motion

The Average Total Density of detected radio frequencies was calculated to have a negative correlation with bee total motion of -0.18. As shown in Figure 4.4, the Average Total Density exhibits a diurnal cycle where it is generally higher at night than during the day. Be motion exhibits an inverse relationship to this pattern.

This trend of higher RF density at night is likely due to ionospheric reflection. The ionosphere is a layer of Earth's atmosphere that becomes ionized when the sun's radiation interacts with it [55]. This ionization during the day causes shifts in the layers of the ionosphere that absorb many communication signals. At night, the layers of the ionosphere change such that more signals are reflected back to earth [56]. This is likely the reason why the observed Average Total Density of RF signals is higher at night and lower during the day, and these regular cycles coinciding with the normal trends of bee activity likely contributed to the strength of the correlation. These patterns of the Average Total Density may be valuable contributors in predicting bee motion.

Avg. RF Watts (W)

The Average RF Watts of detected radio frequencies in the range of 240 MHz to

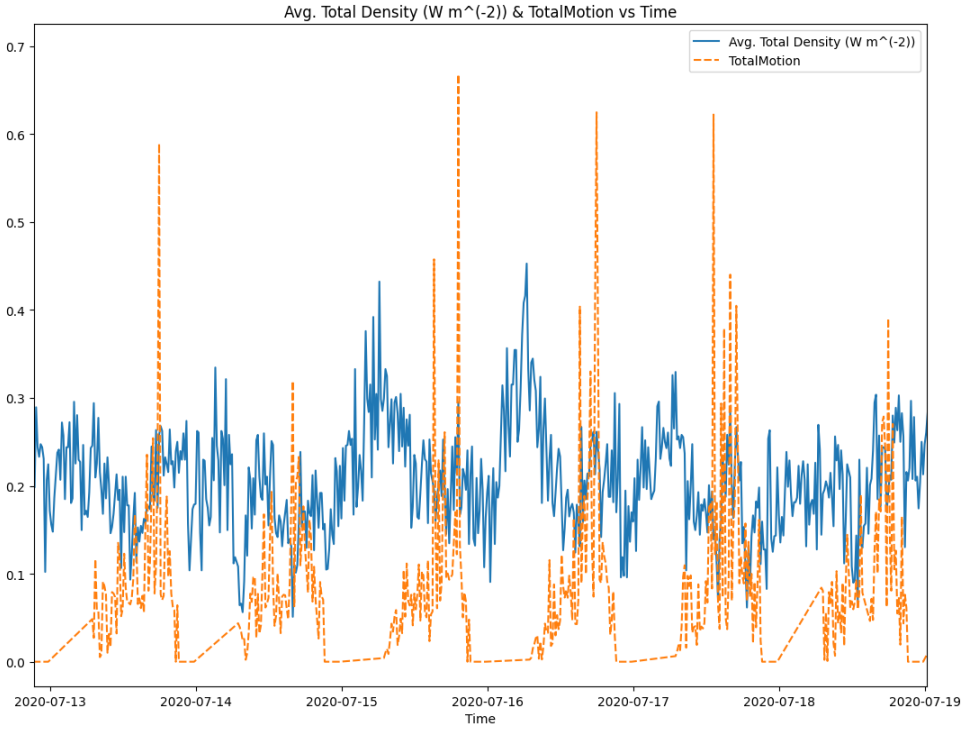

Fig. 4.4: Normalized Avg. Total RF Density and Total Motion plotted against time. Note that the average density follows a diurnal cycle where it is generally higher at night than during the day. This trend is likely due to RF signals being able to travel better at night due to how the properties of the ionosphere change when the sun's radiation isn't interacting with it.

1040 MHz indicated a slight negative correlation of -0.053 to bee total motion. Since this frequency range is a subset of those represented by the Average Total Density, these values followed similar patterns described in the previous section, although slightly less pronounced. That is, it tended to be higher at night than during the day, as shown in Figure 9. Thus, it may also provide information that could contribute to a predictive model.

Avg. EMF and Total Motion

The Average EMF readings had a strong positive correlation of 0.44 with bee total

motion. Figure 10 shows the EMF values rising each day and falling each night, along with the bee motion and temperature. However, the correlation value of average EMF with Temperature was reported as 0.93, which seemed abnormally high.

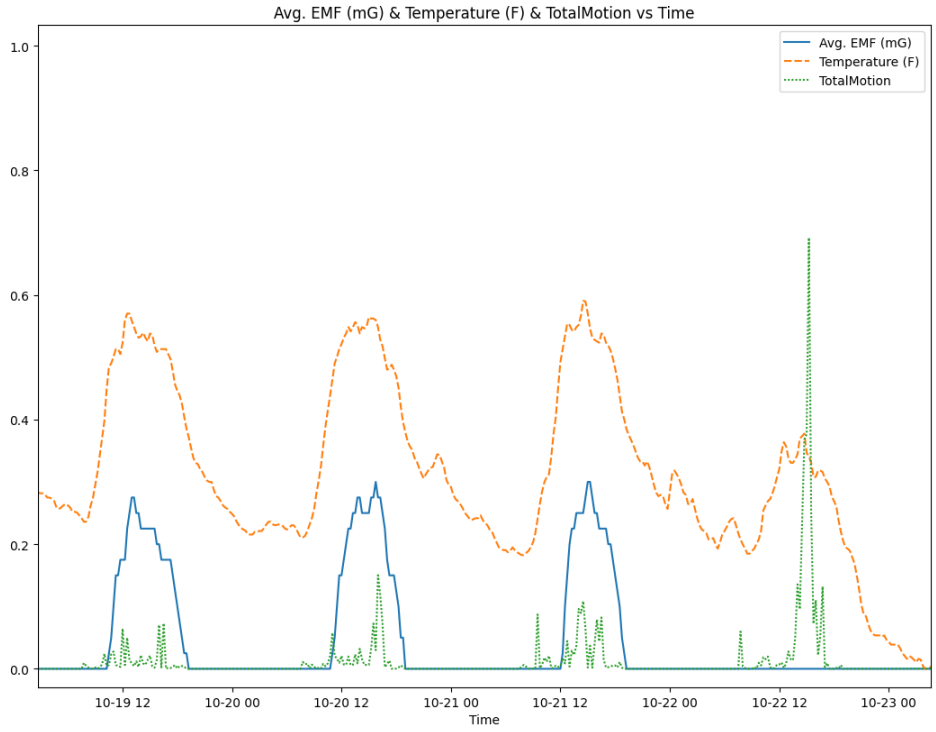

Fig. 4.5: Normalized Avg. EMF, Temperature, and Total Motion plotted against time. Note that the EMF and bee motion rise and fall together for all days except October 22nd. On this day, the EMF value didn't rise above zero where the temperature didn't go above 52.5°F. Similarly, EMF appears to be absent during each night when the temperature was below that level as well.

Upon further investigation of the relationship between temperature and average EMF, it was observed that the EMF values usually remained at zero until the temperature rose above approximately 60°F in the morning, or fell below approximately 53°F in the evening. While the temperature was above the 53-60°F range, the EMF value fluctuated nearly in tandem with the temperature.

This relationship can be seen in Figure 4.5 where the EMF value remained at zero during a day where the non-normalized maximum temperature was 52.5°F. This behavior was observed on both Weather and EMF Sensing Stations, with the caveat that the second station began recording non-zero EMF values starting at approximately 66°F on the rising edge of the temperature, and stopped recording non-zero values at approximately 57°F on the falling edge.

To further verify this behavior, we used a third EMF-390 sensor to collect readings in a lab on Utah State University's campus for a few weeks. The EMF readings in the lab usually fluctuated between 1.4 to 2.1 mG, following a cyclic daily pattern with the low in the early morning. The values never reached zero. EMF readings with the same values recorded by the Weather and EMF Sensing Station readings at the bee hive site corresponded to temperature readings of approximately 65 to 73°F. These temperatures are similar to those maintained in the lab where the tertiary data was collected.

These findings indicate that the EMF-390 sensor does not account and correct for changes in temperature. With this discovery, we decided that the Avg. EMF variable readings as collected by the selected sensor could not be relied upon, and they were not used in the creation of our prediction models.

Avg. EF and Total Motion

The Average EF was calculated to have an overall weak positive correlation of 0.067 with bee total motion. However, it can also be seen in Figure 1 that EF appears to have a strong positive correlation with Relative Humidity, which has a negative correlation with bee total motion. Inspection of Figure 11 shows the general trend of EF being inverse to bee total motion. It can also be seen that it rises and falls in tandem with Relative Humidity. This positive relationship between relative humidity and EF concurs with existing research [57] [58].

Whether these fluctuations are caused purely by changes in relative humidity or not, these patterns my contain additional information for certain situations that could be used by a model to predict bee motion to some degree.

4.1.2 Correlation Comparisons with Other Research Data

Since data collection techniques, climates, bees, and other factors vary from one location to another, we will also present a brief comparison of some of our collected data correlations with some of the correlations reported from the data collected by Clarke and Robert [24]. These comparisons can be found in Table 4.1.

Variable	Our Data	Clarke-Robert	Polatto
Temperature	0.43	0.83	0.62
Shortwave/Solar Radiation	0.34	0.81	0.83
Humidity	-0.35	-0.74	-0.80
Precipitation	-0.038	-0.06	N/A
Atmospheric Pressure	-0.14	-0.08	N/A
Wind	0.19	0.42	0.79

Table 4.1: Comparisons of Pearson correlation coefficient values for overlapping variables between our collected data and the values reported by Clarke and Robert [24] and Polatto et al. [25] for their collected data.

As shown, it's clear that most of the correlations in our data are much weaker than those of the data collected by Clarke and Robert or Polatto et al. There are many factors that could be contributing to these differences. For example, the location and environment where the data was collected is very different. The data collected by Clarke and Robert was in a rural farm area in Southern England (51.4237 latitude, -2.6711 longitude) [24], and the data of Polatto et al. was collected in the Midwest region of Brazil [25]. Compared to our location within the city of Logan, Utah, at a much higher elevation in the Western United States, their locations would be inherent to very different weather and environmental conditions.

Another difference is that the data was collected very differently. In the case of Clarke and Robert, electro-optical bee counters were used to count bee egress rate [24]. Polatto et al. manually captured bees periodically at various foraging sites and counted them [25]. These methods of estimating foraging activity may have very different accuracies than the DPIV method we used to obtain bee motion counts. However, inserting electro-optical bee counters at the hive entrance or manually capturing bees is contrary to our design principle

of preserving the bee space. Our method of video monitoring doesn't interfere with honey bee phenology.

Regardless of the differences, it's important to note that since our study is using different data with lower correlation scores, we can expect to have different prediction results - even if using the same method of prediction. To our knowledge, the data collected by Clarke and Robert or Polatto et al. is not available to the public.

4.1.3 Additional Insights

This analysis has shown that many of the collected variables exhibit distinctive features that could be utilized to predict bee motion in certain conditions. However, it has also shown that one variable alone likely won't be able to accurately predict bee motion throughout the day during varying environmental conditions. For instance, while the temperature may stay constant, rain, wind, or lower shortwave radiation values may cause bee motion to decrease. Thus, each variable when used together will contribute unique valuable information that could be used in a predictive model.

Additionally, this analysis has identified that the Shortwave Radiation and EMF values as collected by our station shouldn't be used in our experiments. These values were significantly affected by other environmental factors such that their use would skew the results of any experiments that involved them.

By exploring the various connections between the collected variables, we have gained insight into what factors may be influencing bee activity throughout the day. While these correlations don't necessarily represent causations, they do provide valuable information that could be useful in future exploration to determine to what the bees *are* specifically responding.

CHAPTER 5

EXPERIMENTS AND RESULTS

After analyzing the data, we proceeded to develop, train, and test a machine learning model that could effectively predict bee total motion. Since we had many columns, some of which are collinear, we had to carefully select which ones should be used to achieve high accuracy while preventing over-fitting. Additionally, we needed to see if performing feature engineering could help improve the overall accuracy. This chapter discusses the process we performed to accomplish these goals, along with the results we obtained.

5.1 Random Forest Regressor

We chose to use a random forest [59] regressor as our machine learning model to predict bee total motion. This algorithm was beneficial since random forests can predict continuous values, typically achieve high accuracy, are fast to train, and they can provide insight into how the variables are used in making each decision. Additionally, random forest models can be built so they don't require large amounts of memory or extensive hardware, like some deep learning models. For the benefit of the reader, we will present a basic overview of how decision trees and random forests are built and used to make accurate predictions.

A random forest regressor consists of multiple decision trees that operate in concert to produce a prediction based on a given input or inputs. Once each individual decision tree has produced a prediction, all predictions in the ensemble are taken into consideration. The final output of the random forest regressor is the average of each decision tree's probabilistic prediction. This is how it can predict continuous values, as opposed to a random forest classifier, which is configured to predict specific classes.

We used scikit-learn's implementation of a random forest regressor [60] [61], which uses an optimized version of the Classification and Regression Tree (CART) algorithm for creating each decision tree in the random forest [62]. As the name implies, this method

works well with either classification or regression data.

A decision tree consists of nodes, edges, and leaves. Each node consists of a test upon an input, which branches to two other nodes or leaves depending on the result of said test.

To create a decision tree for regression, a root node, based on one of the input features, needs to be selected first. This is done by considering each input feature to determine which one will best split the data. In the case of regression, the data of each feature also needs to be optimally split at each node. This root node selection is done by calculating a split value for each feature such that the mean squared error (MSE) of each respective feature's data is minimized. This minimization value is the mean value of the samples of the feature in question.

Once the split value and minimized MSE is calculated for each feature, the feature with the smallest MSE is selected as the root node, with its associated optimum split value. This process continues recursively with the remaining features and the pertinent partition of the preceding node's feature data until a stopping condition is met. Such a stopping condition could entail growing the tree until each feature can't be split any further, or it could terminate when the tree reaches a certain depth, for example.

After the tree has been created, a set of inputs can be fed through it to obtain a prediction. As the data is fed through the tree, the input data is checked to determine whether the node's test condition is met. The result of each test dictates which node will be traversed next, until a leaf is reached. Once a leaf is reached, the value of the leaf is provided as the prediction. If a leaf consists of multiple values, the average of the leaf's values is provided instead.

Alone, decision trees can over-fit the data - especially if the tree is allowed to fully grow until a single output value pertains to each leaf. This prevents it from generalizing well to data inputs that weren't present in the training set used to create the tree. Various techniques are used to prevent this, such as pruning or enforcing various early stopping conditions while creating the tree.

Another technique for preventing over-fitting is using an ensemble of decision trees, as

is done in random forests. When the decision trees of a random forest are built, the features used for each tree can be a random subset of the original features. Also, bootstrap samples can be used when building the trees. This essentially means that a new data set is created by randomly sampling the original data while allowing repeat selections. Since each tree may have a different combination of input features as well as different training data sets, each tree will be better or worse at making some predictions than others. When inputs are fed through each individual tree, the outputs from each are averaged to produce the final prediction. This results in a generalized model that is more robust to over-fitting, and typically performs better than any single decision tree alone.

Decision trees also have the ability to provide measurements of how important each feature is in making predictions. Since the root node of the tree was selected by its ability to best split the data, it contributes to the decisions of more input variables, and has more decisional importance than another node lower in the tree. Thus, the expected fraction of the samples each node is expected to contribute to can be used to estimate the relative importance of the features [61]. In the case of a random forest, the predictive ability of several decision trees can be averaged to produce a feature importance ranking and contribution amount of each feature.

5.2 Model Creation and Selection

We executed a tiered approach to creating a model that could predict bee total motion with a high degree of accuracy. Due to the large number of features, it was first necessary to determine which ones were most important in predicting bee total motion, and eliminate those that didn't contribute significantly. After determining the most important base features to include, we incorporated additional columns that represented the trend of each selected base column. These were added since it was possible that the bees were responding to the changes in the values, rather than the values themselves. Following the addition of the trend columns, we fine-tuned our random forest regressor model's hyper-parameters. Finally, we analyzed the model's predictive ability over different data periodicities.

We used version 0.24.2 of the scikit-learn Random Forest Regressor library [60] and

Python version 3.8.10. To perform each of the tests up until we fine-tuned the random forest regressor model, we configured the random forests to use 100 estimators, and allowed a maximum tree depth of six to keep the trees reasonably sized. We also used the MSE to measure the quality of a split, enabled all features when looking for the best split, allowed bootstrap samples when building the trees, set the random state to 1234 for consistent testing across models, and enabled parallel processing. The remaining hyper-parameters were left at the default values.

The tests were performed using the Ubuntu 20.04 x86-64 operating system with a six core, 3.10 GHz, Intel Core i5-8600 processor, and 15.5 GB of memory.

For each model, we calculated the R-squared value (coefficient of determination), the corrected Akaike information criterion (AICc) value, the weight of the model's AICc value in relation to all others, the 95% confidence interval for the predictions, and each feature used in the model along with the associated importance values for each model's performance on the test data set. These values were used to provide metrics whereby the models could be accurately compared in order to select the one that performed the best.

The R-squared value (unadjusted) is used to indicate the goodness of fit of the model, or how well the model outputs could be predicted based on new input samples [63].

The AICc value is used to estimate the quality of a model in comparison to other models trained on the same data set. It rewards a model for having a good fit, but at the same time penalizes a model for having additional features. By doing this, it discourages over-fitting that can be caused by adding too many features, or features that are co-linear. A model with a lower AICc score than another trained and tested on the same data set is deemed the better model [64]. An AICc value can't be used alone to assess the performance of a model; it can only be used in comparison against resulting AICc values produced by other models on the same data. As will be described below, we trained thousands of models using different combinations and numbers of features on the same data set for each model, and selected the model that produced the lowest AICc score. This effectively resulted in a feature elimination process by selecting the model with the best fit, limited over-fitting,

and the fewest features.

The weight of the AICc value in relation to all other models' AICc scores is calculated by sorting the all of the models by AICc value from least to greatest. Then the delta is calculated between the lowest AICc score and every other AICc score. Once the delta is calculated for each model, the weight is calculated. This weight value represents the probability that a model is the best model in comparison to the others in consideration. The equation used to calculate the weight is given in 5.1, where W_i is the weight of a given model, Δ_i is the difference between the minimum AICc score of all models and the model in question, and n is the number of models being compared. The sum of all calculated weights equals one. A model that has a calculated weight near one exhibits a high probability that it would be selected as the top model if the test was repeated.

$$W_i = \frac{e^{\frac{-1}{2}\Delta_i}}{\sum_{j=0}^{n} e^{\frac{-1}{2}\Delta_j}} \tag{5.1}$$

By measuring the 95% confidence interval for the predictions, we can determine how close the predictions are in relation to the actual values 95% of the time. The smaller the confidence interval, the better the model.

In capturing the features used in a model with their associated feature importance values, we are able to gain some insight into how the features were used by the model to predict bee total motion. It also gives insight into which features appear to be the most important and by how much.

The R_4_5_s1_2020.csv data file was used for the experiments and results in this chapter unless otherwise noted. The scikit-learn "train_test_split" function was used to select a randomly sampled 80% of the data for training, and the remaining 20% of the data for testing the trained model. A single training and testing data split was used for all models trained and tested to ensure that the same respective data samples were used for each test for accurate comparisons. The test data was always kept separate from the training data to avoid skewing the results. All of the tests, besides the time grouping tests, predicted bee total motion at 15 minute intervals, which is the same periodicity as the input data.

It should also be noted that some additional data preprocessing was done immediately before training the random forest models. We replaced the NaN values in the total motion column with zeros since most NaN values occurred during night hours when there wasn't any foraging activity. We also linearly interpolated missing row values for up to four consecutive missing records. If there were gaps larger than this, they were deleted from the data set. Additionally, we normalized the data by min-max scaling so all values would lie between 0 and 1. This was done with the intent that a model trained on data with certain ranges could apply to another hive with a different bee population. It essentially allows the model to focus more on percent values of the total feature ranges, rather than hard-set values. Finally, we also added a month and an hour column to the data so time and seasonality could be taken into account by the model.

Base Feature Selection

As mentioned before, the first step of developing our model required us to select which columns (or features) of the original data should be used. Since our data contained approximately 28 columns, an exhaustive search consisting of every possible combination of these columns was impractical and would simply require far too much time.

As a result, we chose to perform a search on the primary columns from each metric. These primary features mainly omitted those that measured peak values. We also removed the frequency columns since they typically didn't exhibit substantial variation, and the data was very noisy. Since the date and time would likely be valuable in bee motion predictions, we extracted the month and hour values from each time stamp, and created a Month and an Hour column. This would theoretically allow the model to account for temporal and seasonal trends. The final columns used in our base feature selection process include the following: 1) Temperature (F); 2) Pressure (mbar); 3) Humidity (%); 4) Wind Direction (Degrees); 5) Wind Speed (MPH); 6) Wind Gust (MPH); 7) Precipitation (Inches); 8) Avg. RF Watts (W); 9) Avg. RF Density (W m^(-2)); 10) Avg. Total Density (W m^(-2)); 11) Avg. EF (V/m); 12) Shortwave Radiation USU (MJ/m^2); 13) Month; 14) Hour.

The feature selection was performed by training a series of random forest regressor

models, where each model consisted of one of the various combinations of the input columns to predict bee total motion. The data and statistical analysis variables described in the previous section were captured after evaluating each model using the test data. These results for each model were written to a single file.

In total, 16,383 different models were trained and tested. This took approximately one hour and fifteen minutes to run on the hardware described in the previous section, with the current version of our software. Note that while the training and testing of the models was parallelized, additional optimizations to the surrounding code could be implemented in the future that could speed up this process.

After the results for every model were captured, the AICc deltas and weights were calculated for each model. This allowed us to select which combination of features produced the best performing model to predict bee total motion. It also provided us insight into how likely it was that the selected model, compared to the rest, was the most optimal. Additionally, we could see the feature importance ranking of the top performing model to gain insight into how heavily the selected features were being used.

Table 5.1 shows the results of the selected top performing model. This model utilized the Temperature, Shortwave Radiation USU, Humidity, Hour, Pressure, Month, and Wind Speed columns. The Wind Direction, Wind Gust, Precipitation, Avg. RF Watts, Avg. RF Density, Avg. Total Density, and Avg. EF columns were thus not used in the top performing model. Since we used the AICc measurement for model selection, these results indicate that the addition of these columns in any combination did not improve the fit of the model while minimizing the variance more than the above-selected model.

The intent of selecting the optimal combination of features to use from the original data was to provide a foundation upon which other models could be compared against it that utilized additional engineered features.

Trend Column Addition and Selection

After selecting the base features from the original data that produced the best performing model, we proceeded to engineer some additional columns that could potentially present

Evaluation Items	Value
R^2 Score	0.533218750292152
AICc Value	-16236.0253821285
Weight	0.999990598187116
Normalized 95% Confidence Interval	(5.373905514464679e-05, 0.004319988214206403)
Non-normalized 95% Confidence Interval Motion Count Span	58.110579796
Features & Importances	0.40957 - Temperature 0.23231 - Shortwave Radiation USU 0.08281 - Humidity 0.07775 - Hour 0.07534 - Pressure 0.06846 - Month 0.05375 - Wind Speed

Table 5.1: The statistical analysis results of the top performing model resulting from the base column feature selection. Note the very high weight value, indicating a high probability that this model would be selected again should the experiment be repeated.

important information to the model that the original columns alone could not. Since bees may be responding to *changes* in conditions rather than the current conditions themselves, we wanted to present this information to the model as well. This feature engineering involved creating trend columns for each of the selected columns over trend intervals ranging from one hour to 24 hours in one-hour increments. This created 24 new columns for each of the seven selected features.

The values for each trend column were calculated by taking each value pertaining to an attribute at each time point and subtracting the value at an earlier time point as defined by the trend interval amount. Let T_{now} represent the current time and $T_{interval}$ represent the trend interval amount. T_{past} is given by T_{now} - $T_{interval}$. The trend value $Trend$ for a new trend column is calculated by taking the value of the attribute at T_{now} and subtracting the value of the attribute at T_{past}. For example, if the temperature at the current time is 75°F, the trend interval is two hours, and the temperature two hours before was 80°F, then the trend value would be -5°F. This was repeated for each column where $T_{interval}$ ranged from 1 hour to 24 hours in 1 hour increments. Each new column was identified by the name of

the original column with the trend interval appended to it (e.g. Temperature (F) 1:00:00 Trend, Temperature (F) 2:00:00 Trend).

With these new trend intervals for each feature, we needed to determine which trend interval for a given feature was most valuable. To do this, we trained a new random forest regressor model using the previously selected base columns and one of the feature's 24 trend columns at a time. We captured the statistical analysis results for every model trained for that feature, and used the lowest AICc value to select the model that performed the best. This exercise was performed for each feature and its associated 24 trend intervals.

After selecting the most useful trend interval for each feature, a search was performed to determine which, if any, of the features' selected trend intervals improved the model. This was done by training a new model for each individual combination of the trend columns in addition to the fixed base columns, and capturing the statistical results for each. In other words, we used the base columns for every model trained, with the addition of a different combination of the selected trend columns for each test.

Once the results for each model were captured, the one with the lowest AICc value was selected as the top performing model. It should be noted that the results for the base column model were included in the comparison to determine if adding trend columns did indeed improve the model. Table 5.2 shows the results of the selected model, which included several trend columns.

As can be seen in Table 5.2, the addition of the nine-hour humidity trend, 8-hour temperature trend, five-hour shortwave radiation trend, and the eight-hour wind speed trend columns improved the model. The R^2 score increased while the confidence interval decreased. This indicates a better model fit. It should also be noted that the very high weight (greater than 99%) of the chosen model indicates that out of all of the models compared, this one has a very high probability of being the top model. It is also noteworthy that the nine-hour humidity trend column manifests itself as the most important feature being used in predicting bee total motion with a feature importance of 0.32521.

Evaluation Items	Value
R^2 Score	0.545481459114656
AICc Value	-16295.9591445764
Weight	0.997348029628673
Normalized 95% Confidence Interval	(-0.0003871414906560602, 0.0038237831215572897)
Non-normalized 95% Confidence Interval Motion Count Span	57.357004143
Features & Importances	0.32521 - Humidity 9-hr Trend 0.13839 - Temperature 0.11091 - Temperature 8-hr Trend 0.11078 - Shortwave Radiation USU 0.06065 - Month 0.05940 - Humidity 0.05020 - Shortwave Radiation 5-hr Trend 0.04862 - Pressure 0.03324 - Wind Speed 8-hr Trend 0.03192 - Hour 0.03069 - Wind Speed

Table 5.2: The statistical analysis results of the top performing model resulting from the base column feature selection and the trend column feature selection. Note the high AICc weight and the narrow confidence interval span.

Random Forest Regressor Hyper-parameter Grid Search

Once the features of the top performing model were selected, we proceeded to fine-tune our random forest regressor model's hyper-parameters. Hyper-parameters are values used by the algorithm that can be adjusted to tune the model's performance. According the scikit-learn's documentation, the two most important hyper-parameters to tune include the number of estimators used, and the maximum depth to which each decision tree is allowed to grow [61]. Thus, we chose to perform a grid search while varying these two parameters of the random forest regressor model and comparing the results of each model evaluated on the test data. The number of estimators we tested ranged from 100 to 500 in increments of 25, and the maximum tree depth varied from 2 to 20 in increments of one. Each test was performed while using the columns described in Table 5.2.

Again, we used the AICc value to evaluate which model performed the best, and the results are shown in Table 5.3. The top performing model was obtained while using 300

estimators and a maximum tree depth of 12. While these changes didn't improve the model significantly, it did improve the R^2 score and narrow the confidence interval slightly. One will note that the weight assigned to the top performing model is lower at 0.47. This is likely due to the fact that the same columns were used for each model. However, the next closest weight is 0.25, so it's still fairly likely that the model selected is the best one out of all those that were compared. Also, some of the feature importance values changed slightly, but this is likely due to the fact that some of the lower-ranked features had feature importance values there were relatively similar in the previous tests.

Evaluation Items	**Value**
R^2 Score	0.546560224896903
AICc Value	-16302.7575852117
Weight	0.470065585641744
Normalized 95% Confidence Interval	(-0.0007492730731365131, 0.0034573617069622612)
Non-normalized 95% Confidence Interval Motion Count Span	57.29857234
Features & Importances	0.32521 - Humidity 9-hr Trend 0.13839 - Temperature 0.11078 - Shortwave Radiation USU 0.11091 - Temperature 8-hr Trend 0.05940 - Humidity 0.06065 - Month 0.04862 - Pressure 0.05020 - Shortwave Radiation 5-hr Trend 0.03324 - Wind Speed 8-hr Trend 0.03069 - Wind Speed 0.03192 - Hour

Table 5.3: The statistical analysis results of the final top performing model resulting from our process of selecting the base features, trend columns, and the random forest regressor hyper-parameters. Note that the 95% confidence interval span is very small, 57, compared to the total range of bee motions, 0 to 13,621.

It should be noted that due to the random nature of the random forest regressor model and the large volume of similar input features, it's possible that subsequent executions of the

process used to arrive at this model may render different results with similarly performing models. This could include slightly different column combinations, feature importance rankings, or even performance. However, this characteristic is also true of other feature selection methods accepted in practice, such as Recursive Feature Elimination and Sequential Feature Selection. The next chapter compares the results of these other feature selection methods to the process we used. Each of those methods also produce slightly different results on subsequent runs. Regardless, the intent of the process described here is to demonstrate a method that consistently results in a performant model that can effectively predict honey bee motion from one moment to the next. The remainder of this thesis will refer to the model whose results are shown in Table 5.3 as our selected top model trained on the original 15-minute data.

Other Feature Selection Methods

When it comes to eliminating or selecting features from data sets that contain many, there are several other methods that are commonly practiced. As mentioned in the previous section, two such methods include Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS). For completeness, we employed both of these methods (using scikit-learn's libraries [65] [66]) on our data set with every original feature along with each hour trend column from one to 24 hours for each column. It should be noted that we removed the Avg. EMF, Max EMF, UpwardMotion, DownwardMotion, and LateralMotion columns, and we didn't calculate trend for the Month and Hour in these experiments. As noted in Chapter 4, we removed the EMF variables due to its apparent skew due to temperature, and the other motion columns should not be treated as inputs since collectively they constitute the TotalMotion variable.

RFE is an algorithm in which a model, such as a random forest, is trained on a data set, and the importances of the features used in making the predictions are captured. Some of the features that have low importances are removed from the dataset, and the process is repeated on the remaining features until the desired number of features remain. These remaining features are deemed the most important contributors in predicting the output.

Forward SFS is a similar algorithm, however, it operates additively. Rather than recursively eliminating features, it iteratively adds features. The algorithm starts with zero features, and then determines which feature when added to the model improves it the most. After a feature is added, it considers the remaining features and likewise determines which one best contributes to the model's performance. This process continues until the desired number of features have been selected.

These methods are desirable because they can narrow down the number of features without having to perform an exhaustive search of every possible combination of every feature. They can also narrow down the number of features in a relatively small amount of time, and do so in a largely automated fashion. Since our data set has many variables, adding the trend columns for each feature inflated our data set to contain 603 different features.

For each selection method, we used a random forest regressor with the model hyper-parameters discussed in the previous section. We performed RFE with the algorithm configured to eliminate 5% of the features after each iteration until it arrived at 11 features (the same number of features arrived at by our original selection method based on AICc scores).

It should be noted that we initially attempted to use the available scikit-learn Recursive Feature Elimination with Cross-validation (RFECV) library [65] to automatically select the features as well as the optimal number of features to include. However, this resulted in the model automatically selecting 92 different features out of the initial 603, and the R^2 score reached about 0.50 with a 95% confidence interval span of about 60 bee motions. Since this didn't produce a better model or sufficiently eliminate unwanted features, we chose to limit it to the same number of features obtained by our AICc feature elimination method, and use the RFE module provided by scikit-learn.

We performed forward SFS with the algorithm configured to use five-fold cross-validation in evaluating each feature until it also arrived at 11 features. For each algorithm's feature

selection process, the 80% test split that was used in all other experiments previously described was used here as well to avoid skewing the test results.

After the 11 features were selected by each algorithm, a final random forest model was trained on the training split and tested on the same 20% test split used in all other experiments to evaluate the overall performance of a model with the selected features. This allowed us to compare the selected features, the feature importance rankings, model accuracies, and 95% confidence intervals to determine which selected model appeared to be the best.

Figure 5.4 shows a comparison between the performance of the models selected by RFE, SFS, and our tiered selection approach. As can be seen, all three methods produced models with acceptable performance. However, the model produced by our approach performed slightly better with both a higher R^2 score and a slightly narrower 95% confidence interval span than the models produced by either RFE or SFS feature selection.

	Our Approach	**RFE**	**SFS**
R^2 Score	0.547	0.525	0.507
AICc Value	-16302.8	-16170.1	-16061.0
Non-normalized 95% Confidence Interval Motion Count Span	57.298	58.635	59.771

Table 5.4: A comparison of the top models produced by our tiered feature selection method, Recursive Feature Elimination, and Sequential Feature Selection as trained and tested on data from the R_4_5 hive at 15 minute periodicities.

Figure 5.5 shows a comparison of the features that were selected by each method along with their associated importance rankings. Our method limited the selected number of trend columns for each feature to only one to help reduce covariance. However, it can be seen that RFE and SFS did not intentionally prevent this since both selection methods selected more than one trend column for multiple variables. Also, the majority of the columns selected by RFE and SFS were trend columns.

Another item of note is that our feature selection approach and that of SFS both selected humidity trend as the most important predictor of bee total motion. While RFE didn't select humidity trend at all, it should be noted that Avg. EF was ranked third in importance, and that this variable was previously shown in Chapter 4 to have a strong positive correlation with humidity. Generally speaking, all three methods selected some form of temperature, shortwave radiation, and month, while the majority of the methods additionally included forms of humidity trend, hour, pressure, and EF.

Time Grouping Performance Analysis

After arriving to our best performing random forest regressor model, we proceeded to evaluate it on different data resolutions. In other words, we wanted to find out how the model performs when the original data records spaced 15 minutes apart are averaged into larger data periodicities before being fed into our model.

Since the original data existed in 15 minute intervals, data spans larger than 15 minutes were averaged together to create a representative point (time grouping or data periodicity) over that time span. The trend variables are still calculated using the original data spaced at 15 minute intervals before each column is averaged over each desired time grouping interval. Time interval groupings were created from 15 minutes up to 24 hours in 15 minute increments (e.g. 0.25, 0.5, 0.75, 1.0, ..., 24.0). A random forest model for each data resolution was trained on 80 percent of the grouped data, and tested on a separate 20 percent of the data.

As can be seen in Figure 5.1, the highest R^2 score obtained was 0.886 with a data periodicity of 12 hours. The confidence interval span, although not the lowest, was decent at 0.0068, which equates to approximately 92 un-normalized bee motion counts. This is about 0.68% of the total bee motion count range of zero to 13,621. In other words, 95% of the time, the predictions were within about 92 bee motion counts of the actual value.

It should also be noted that the original data periodicity of 15 minutes resulted in a smaller R^2 score of 0.542, but it also had the smallest confidence interval of 0.0042, or 58 un-normalized bee motion counts. So, although the R^2 score was lower, it's predictions

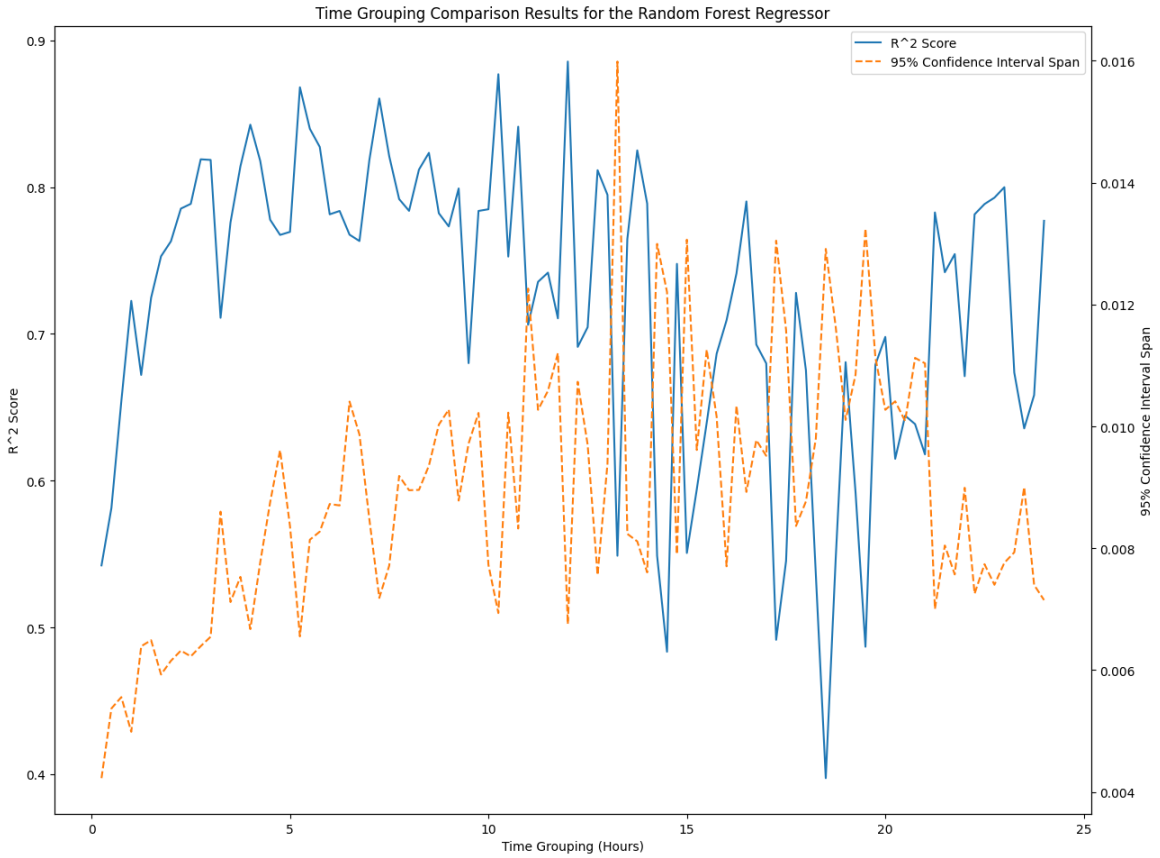

Fig. 5.1: The R^2 score and the normalized data's 95% confidence interval span results for each random forest regressor model trained on a different time grouping. A higher R^2 score indicates a better fit, while a lower 95% confidence interval span shows that more of the predictions were closer to the actual values.

were far closer to the actual values most of the time.

Another item of interest is that there are several other data periodicities in between 15 minutes and 12 hours that also produced satisfactory results. For example, at 2.75 hours the R^2 score was 0.819 with a 95% confidence interval of 0.0064, or about 87 bee motion counts. Additionally, at 5.25 hours the R^2 score was 0.868 with a 95% confidence interval of 0.0066, or about 89 bee motion counts. However, at periodicities over approximately 10.75 hours, the results become increasingly unstable. Periodicities below this have R^2 scores and 95% confidence intervals that are much more stable relative to each other.

It should also be noted that when the data was grouped into larger segments, there were fewer data points used in both the training and the testing data sets.

Overall, these results show that relatively high accuracy and 95% confidence intervals can be obtained at a periodicity low enough to be sufficient for bee hive monitoring and timely alerting.

5.2.1 Learning Transferability

Since all of the inputs used to predict the bee total motion came from the Weather and EMR Sensing station, a model trained for one hive will have a limited ability to predict the bee total motion of another hive. This shortcoming is exacerbated by the fact that we have six hives at the same location where there is one Weather and EMR Sensing Station being used to monitor the weather and EMR for all of the hives. A model trained to predict the bee total motion based on the weather and EMR may do very well on a single hive. However, if a model pre-trained on one hive is used to predict the total motion of another hive, it will produce the same predictions that it did on the first hive since it received the same inputs.

This deficiency could likely be remedied by adding one or more inputs that come directly from the hive being monitored. Such inputs could include hive buzz intensity or in-hive climatic variables such as temperature or CO_2 levels, for example. If some of these variables were added as inputs, the model would likely be able to better generalize to other hives experiencing the same weather and EMR conditions since the input would be slightly different for every hive. Additional research would be needed to verify this.

However, since the data was normalized before it was used to train and test our model, it should have some ability to compensate for things such as hive population differences. If one hive had a much larger population than the hive the model was trained on, assuming the health status of the hives was similar, the model trained on the smaller hive should be able to predict the total motion of the larger hive. This is because the predictions are basically represented as percentages of the max bee total motion detected over any 15 minute time interval, rather than hard-coded values.

Regardless of the transferability limitations, we proceeded to test the model's ability to transfer learning from one hive to another for completeness. To do this, we trained a

model on the R_4_5 hive using the R_4_5_s1_2020.csv file with 15 minute increments, and then saved it. We then used this trained model to predict the bee total motion values of the R_4_11 hive (data contained in the R_4_11_s1_2020.csv file), and obtain the accuracy results. We chose this hive since it, like the R_4_5 BeePi monitor, didn't experience any hardware failures, and the data spanned the same time range as the R_4_5 hive.

After predicting the bee total motion for the R_4_11 hive, we calculated the R^2 score for the predictions. This resulted in an R^2 score of about 0.39. These results are lower in comparison to the results obtained when using the test set of the R_4_5 data (see 5.3). However, the fact that there is still some predictive ability shows that the hives do show similar trends based on the inputs.

That said, when an 80%-20% data split was performed on the R_4_11 hive's data, and a model was trained and tested for this hive, an R^2 score of about 0.60 was achieved with a 95% confidence interval span of 47 bee motion counts when predicting at a 15 minute periodicity. This is 0.06 higher than the results obtained for the R_4_5 hive. An analysis of the model trained on this hive's data at different periodicities achieved a maximum R^2 score of approximately 0.91 at 20 hours and 15 minutes (also higher than the maximum R^2 score reached for the R_4_5 hive). This shows that while learning doesn't transfer well from one hive to another, the random forest regressor model with the selected columns and hyper-parameters can do well when trained on a specific hive's data. In other words, our developed model configuration could be utilized by beekeepers to train their own specific models for their hives, and achieve approximately the same accuracies we observed in our testing.

5.2.2 Other Model Comparisons

We also used two other algorithms to compare their performance against that of the random forest regressor model. These included the K-Nearest-Neighbors and Partial Least Squares Regression models. We trained each model using the columns that rendered the optimal random forest regressor model with the same train and test splits. For each test, we generated the R^2 Score and the 95% Confidence Interval to be used in evaluating the

performance of each model against the random forest regressor.

Partial Least Squares Regression

Since other research utilized partial least squares regression to predict bee activity [24] [26], we decided to use the same algorithm to evaluate how it compares to the random forest regressor algorithm. To do this, we used the PLSRegression module from the scikit-learn library [67].

The provided python module for the PLS regressor allowed the specification of different numbers of components to keep in the algorithm. This number of components is required to be set to a value between one and the minimum value between the number of samples and the number of features. Since the number of features is 11 for our model, we decided to set this value to a maximum of 11.

We thus trained and tested 11 different models where each model utilized a different number of components from one to 11. After capturing the results for each model, we selected the top performing model based on its maximum R^2 score and it's smallest confidence interval. This top performing model was achieved when the number of components was set to 11. With the original 15 minute data, this model reached an R^2 score of 0.470 with a 95% confidence interval span of 0.0040 (equivalent to a non-normalized span of about 54 bee motion counts).

We also performed the same time grouping tests that were done with the random forest regressor while using the PLS algorithm. As can be seen in Figure 5.2, at 15 minute intervals (the original data periodicity), the PLS model achieved a lower R^2 score than the random forest model. However, the 95% confidence interval span was slightly smaller for PLS than the random forest.

It should also be observed that while the PLS and random forest models perform fairly similarly in general, the random forest model maintains a higher average R^2 score of about 0.72 over the time groupings from 15 minutes up to three hours than the PLS model, which has an average R^2 score of about 0.66. The average 95% confidence interval span over the

time groupings from 15 minutes up to three hours for both models is nearly the same with and R^2 score of about 0.0059.

Another observation is that the PLS model appears to achieve better stability as the time groupings increase. In fact, the model's highest R^2 score value of about 0.91 and smallest confidence interval span of about 0.0039 occurred at a time grouping of 20.25 hours. While these values are slightly better than the random forest regressor model, there is much less training and testing data points being used at this time grouping. This limits the backing for these results. Additionally, it is more valuable to have a higher accuracy at a lower data periodicity in order to provide more timely beekeeper alerts.

The main takeaway from this comparison is that the random forest regressor performs very similarly to PLS regression, and it can perform slightly better at lower periodicities.

K-Nearest-Neighbors

We also chose to train and test a K-Nearest-Neighbors regressor model for an additional comparison against our top random forest regressor model. To do this, we used the KNeighborsRegressor module from the scikit-learn library [68]. Using the original 15 minute interval data, we trained and tested 20 different models where each model was trained using a different number of neighbors from one to 20. We captured the test results from each model and selected the one that achieved the highest R^2 score with the smallest confidence interval. This occurred when there were 16 neighbors used.

As shown in Figure 5.3, the KNN model achieved very good results on the original 15 minute interval data. It achieved an R^2 score of about 0.59 with a 95% confidence interval span of about 0.0035 (equivalent to a non-normalized span of about 47 bee motion counts). This was better than the random forest regressor model. Additionally, for the time groupings from 15 minutes to three hours, the average R^2 score of about 0.70 was just slightly below the average for the random forest regressor of about 0.72. The 95% confidence interval span average over those time groupings for the KNN model was slightly smaller at about 0.0056 compared to the random forest regressor model's average of about 0.0059.

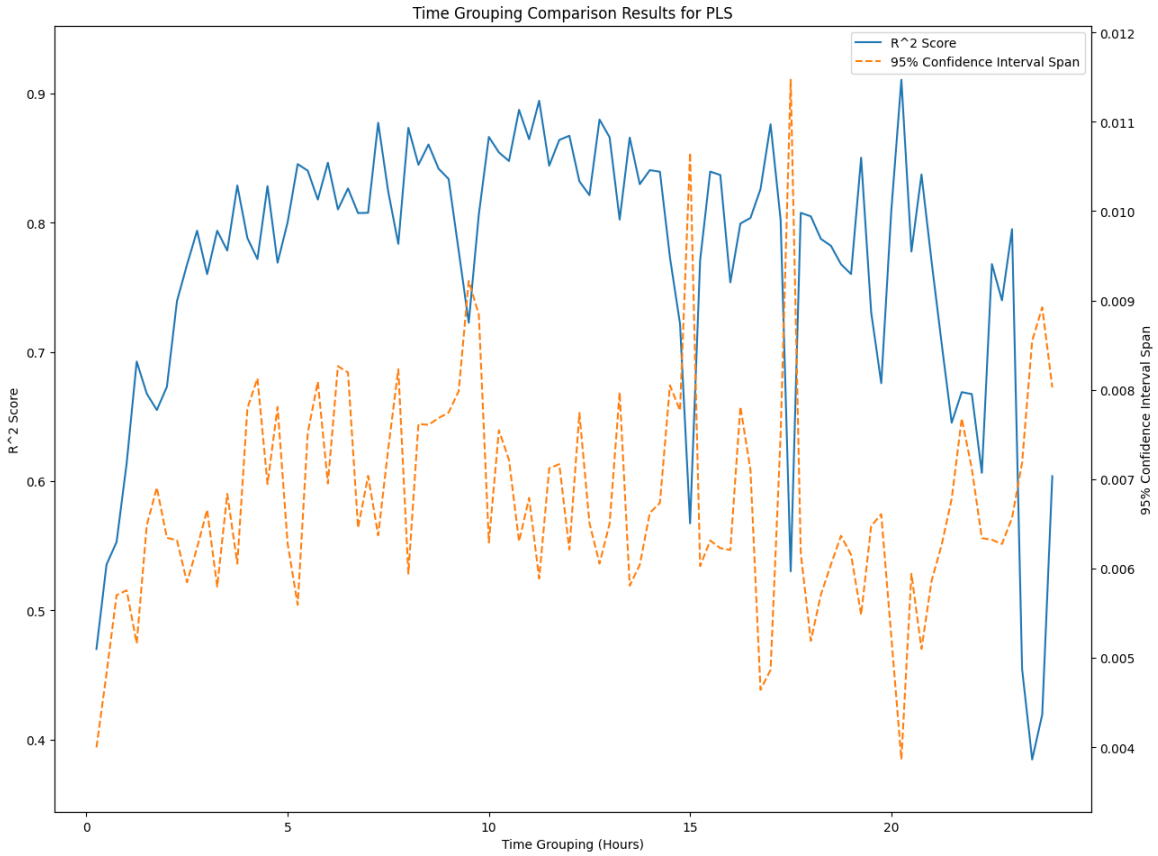

Fig. 5.2: The R^2 score and the normalized data's 95% confidence interval span results for each partial least squares regression model trained on a different time grouping. A higher R^2 score indicates a better fit, while a lower 95% confidence interval span shows that more of the predictions were closer to the actual values.

However, as observed, the KNN model's 95% confidence interval span increases rapidly and the R^2 score decreases rapidly as the time grouping intervals increase. The highest R^2 score is obtained at a data periodicity of six hours where there is an associated 95% confidence interval span of about 0.0063. This peak R^2 score is just a little below that obtained by the random forest regressor model, but with a slightly narrower confidence interval. It also occurs at a smaller data periodicity than was obtained by the random forest regressor.

This comparison shows that the KNN model exhibits a performance similar to that of the random forest regressor at small data periodicities, but worse at larger data periodicities. The random forest regressor also provides the added benefit of being able to view the

importances of each feature used in the predictions.

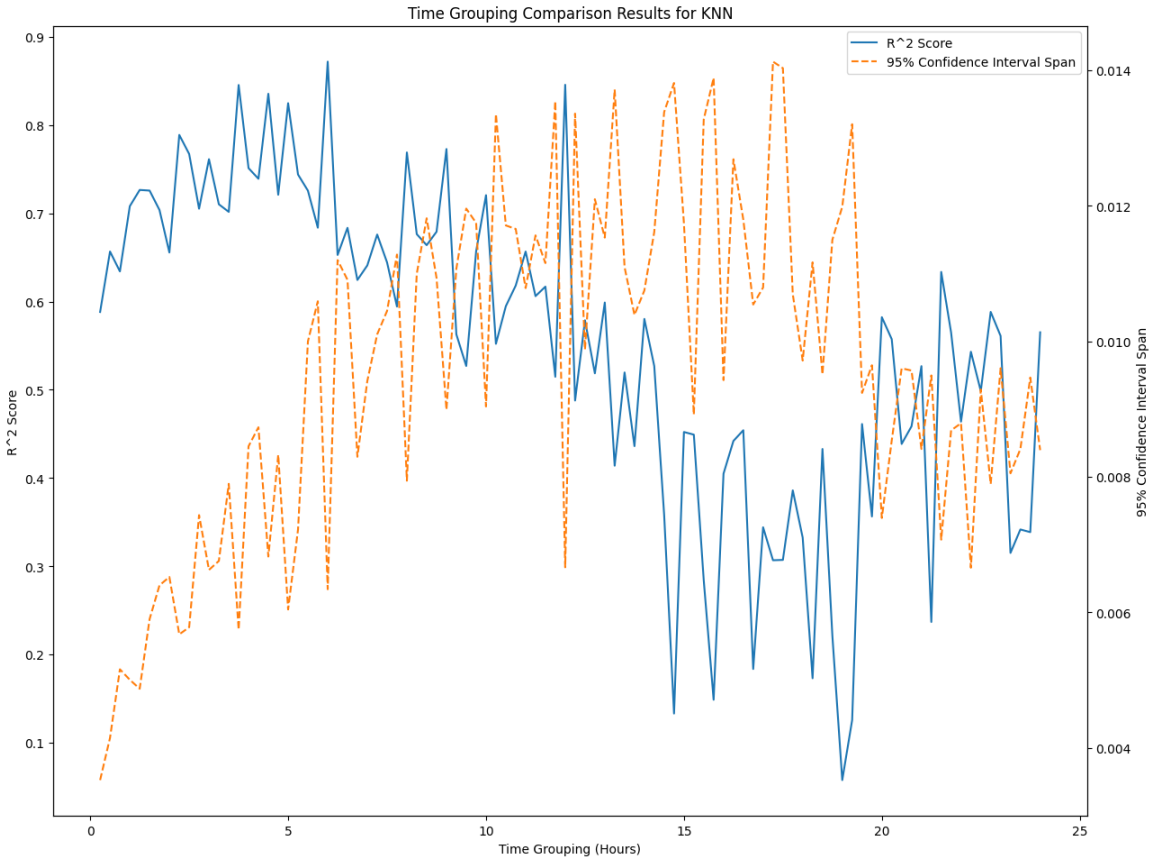

Fig. 5.3: The R^2 score and the normalized data's 95% confidence interval span results for each K-Nearest-Neighbors regression model trained on a different time grouping. A higher R^2 score indicates a better fit, while a lower 95% confidence interval span shows that more of the predictions were closer to the actual values.

5.2.3 Experiment Results Discussion

The experiments in this chapter have shown that a random forest regressor model can be used to predict bee total motion with acceptable accuracy. We've also been able to determine that the features shown in Table 5.3 produced the best model while minimizing the variance introduced by each added variable.

Additionally, by using a random forest and our tiered feature selection approach, we were able to gain insight into which variables contributed the most in predicting the bee

total motion with the highest level of accuracy. Namely, it was discovered that the nine-hour humidity trend was the most important variable in our top selected model, followed by temperature and shortwave radiation. This finding is significant since several other papers have found that temperature and shortwave radiation were the two most significant contributors to accurately predicting bee activity [24] [23] [26].

To delve into this finding further, we looked at the performance results of models trained using each combination of the temperature, shortwave radiation, and the nine-hour relative humidity trend. This allowed us to see how the model performs with each input variable alone, as well as in different combinations of the others. This gives additional insight into the importance of each variable in the prediction of bee total motion.

The R^2 score, AICc value, and non-normalized 95% confidence interval span of each of these models is shown in Figure 5.6. As can be seen, the shortwave radiation when used alone renders a better model than when the humidity trend and the temperature are used alone. However, the humidity trend renders a better model than when the temperature is used alone, but when either humidity or temperature is used in combination with shortwave radiation, similar performance levels are achieved. As far as we are aware, this is the first time the humidity trend has been utilized and found to be a major indicator in predicting bee activity.

We also determined optimal trend intervals for other variables, and found that adding the eight-hour temperature trend, five-hour shortwave radiation trend, and eight-hour wind speed trend improved the model without significantly increasing the variance.

In the process of arriving to the optimal model, we found that it appeared that the ambient EMR variables as monitored by our selected sensor at the data collection site did not significantly contribute to predicting bee total motion at the hive entrance. When these variables were added, they did not improve the model fit while keeping the variance low. To further test this, we trained a random forest model using only the Average Total Density (since it had the strongest correlation with bee total motion) to see how much it could contribute to predicting bee total motion. This resulted in a model that produced an R^2

score of about 0.02542, and a non-normalized 95% confidence interval span of about 84 bee motion counts. These results are significantly lower than the results of the humidity trend, temperature, and shortwave radiation values when used alone to predict the bee total motion. However, it does indicate that the variation in the Average Total Density explains approximately 2.5% of the variation in bee total motion throughout the day.

We also trained and tested a model consisting of just the Average EF and just the Average RF Watts to see how well they could predict bee total motion. The model consisting of EF alone produced an R^2 score of approximately 0.07, and the model consisting of the Average RF Watts alone produced an R^2 score of approximately 0.00. It's reasonable that the EF variable produced a model with measurable accuracy due to its strong correlation with relative humidity, but it appears that the Average RF Watts of the frequency band monitored is not useful in predicting bee total motion.

Thus far, our experiments have shown that our *best* bee motion prediction model is obtained without the use of any EMR variables, we have yet to show how well bee motion can be predicted when using only EMR variables. To test this, we trained and tested another random forest regressor model while only using Avg. RF Watts, Avg. RF Density, Avg. Total Density, and Avg. EF. In doing so, we obtained an R^2 score of 0.1874 with a 95% confidence interval span of approximately 77 bee motion counts. While this isn't as high as our best model, it does show that the variance in these EMR variables can be used to estimate nearly 19% of the variance in bee motion at the hive entrance.

Beyond variable selection, these experiments also characterize the performance of the top model at various data periodicities. These results showed that the random forest regressor model can produce fairly accurate predictions with reasonable confidence intervals from about 15 minutes up to around 12 hours, with several optimal periodicities in between. In comparison to the KNN and PLS regression models, the random forest regressor model performed on par, and in some cases better. We observed that the KNN model achieved its best performance at lower periodicities and the PLS algorithm performed its best at mid-range periodicities. The random forest regressor model performed its best at smallest

periodicities, and fairly well at mid-range and large periodicities. Table 5.7 shows the average R^2 score across all data periodicities, as well as the average 95% confidence interval span across all data periodicities. As can be seen, the partial least squares model achieved the best performance across all periodicities with the exception of the random forest regressor getting a slightly higher R^2 score across the first 8 hour periodicities. Overall, this showed that the random forest regressor exhibits competitive performance against other common models.

As mentioned earlier, since the model did not use any hive-specific variables to predict bee total motion, the learning transferability was very limited. However, it's possible that this limitation could be overcome by adding some hive-specific variables. Also, since data was only collected from one season for these experiments, the collection of additional data would likely improve the accuracy and ability of the model to generalize to other hives.

Finally, by using this method of bee motion prediction, accurate and timely alerts could be given to beekeepers. Figure 5.4 shows the predicted values compared to the actual values on the original data periodicity of 15 minutes. As can be seen, while the overall fit could definitely be improved, the predictions capture some of the prominent changes in bee total motion and general trends. That being said, the predictions are usually not too far from the actual values. Given this information, alerts could be triggered when the actual bee motion count value at a given time is outside of the predicted value's 95% confidence interval bounds, plus or minus some selected buffer to produce useful alerts at such a small data periodicity.

Additionally, Figure 5.5 shows the predicted values compared to the actuals from a model trained on 5.25 hour periodicity data. At this periodicity, the predictions are very close to the actual values, and even more reliable alerts could be produced. While even higher accuracies can be obtained at both 10.75 and 12 hours, a balance must be struck between how early a beekeeper needs to be notified, and what kind of accuracies are required. For instance, if a beekeeper is mainly interested in being alerted to general hive health deterioration, 12 hour or longer monitoring would probably be sufficient. However, if a

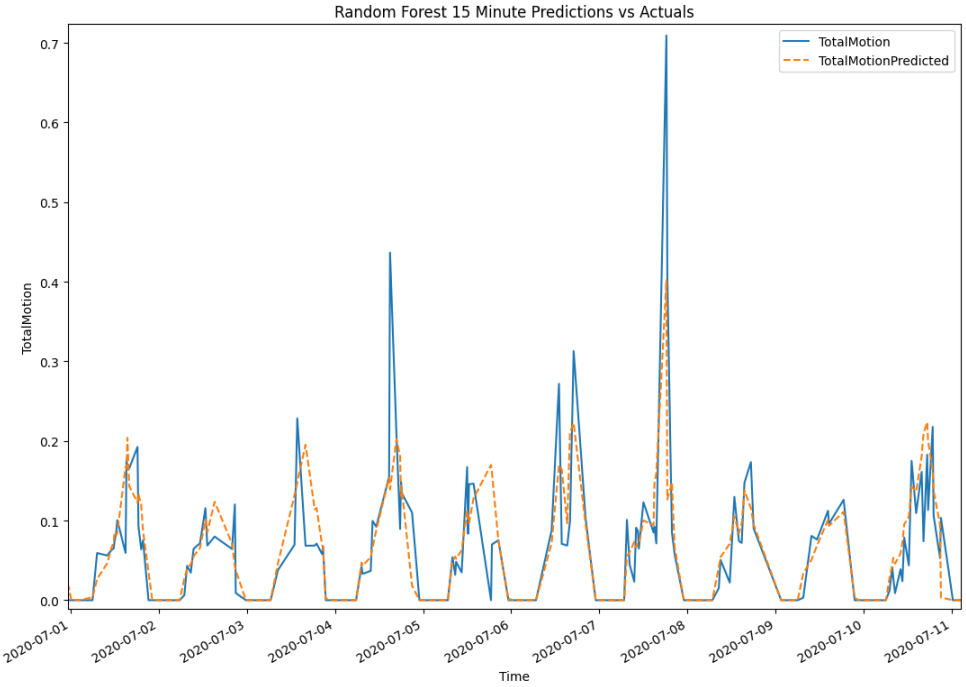

Fig. 5.4: The actual normalized bee total motion values from the test set plotted against the model's bee total motion predictions at the same time points using 15 minute periodicities. Note that some of the prominent features are captured along with the general trend of each day.

beekeeper is more interested in being alerted to swarming events, shorter periodicities, such as 15 minutes, would be more valuable.

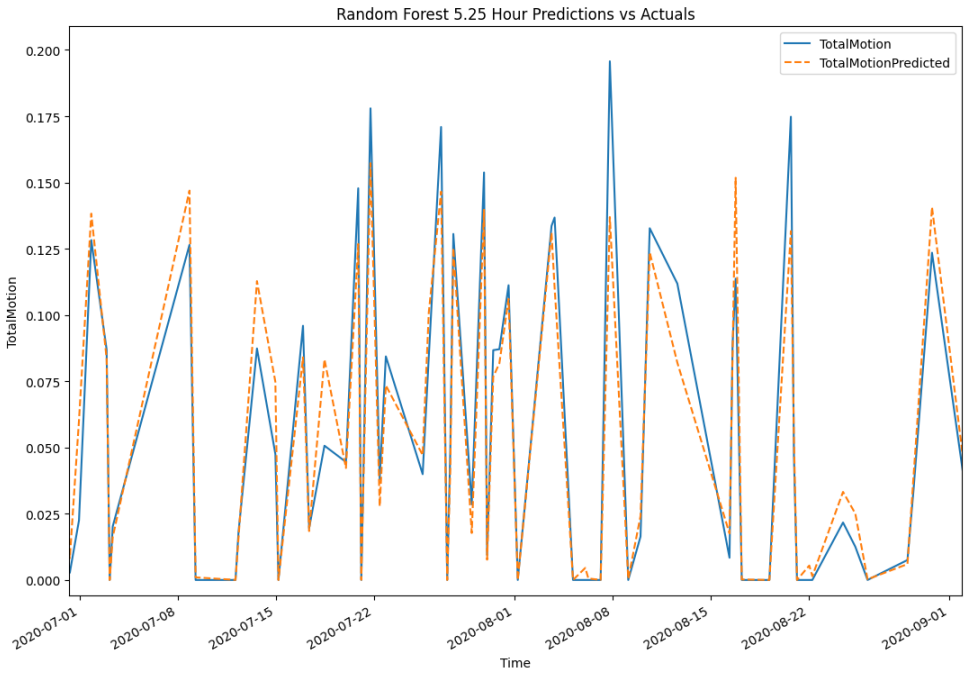

Fig. 5.5: The actual normalized bee total motion values from the test set plotted against the model's bee total motion predictions at those same time points using 5 hour and 15 minute data periodicities. Note that the fit is much better at this data periodicity than at 15 minute data intervals.

Selection Method	Features & Importances
Our Approach	0.32521 - Humidity 9-hr Trend 0.13839 - Temperature 0.11078 - Shortwave Radiation USU 0.11091 - Temperature 8-hr Trend 0.05940 - Humidity 0.06065 - Month 0.04862 - Pressure 0.05020 - Shortwave Radiation 5-hr Trend 0.03324 - Wind Speed 8-hr Trend 0.03069 - Wind Speed 0.03192 - Hour
RFE	0.34413 - Shortwave Radiation USU 19:00:00 Trend 0.11654 - Temperature 0.08372 - Avg. EF 0.07774 - Temperature 9:00:00 Trend 0.07748 - Temperature 2:00:00 Trend 0.05534 - Shortwave Radiation USU 0.05511 - Max Total Density 15:00:00 Trend 0.05320 - Temperature 19:00:00 Trend 0.05245 - Shortwave Radiation USU 20:00:00 Trend 0.04249 - Shortwave Radiation USU 1:00:00 Trend 0.04181 - Month
SFS	0.39399 - Humidity 7:00:00 Trend 0.10992 - Month 0.10976 - Temperature 19:00:00 Trend 0.08539 - Temperature 2:00:00 Trend 0.08250 - Shortwave Radiation USU 6:00:00 Trend 0.06309 - Hour 0.05623 - Pressure 12:00:00 Trend 0.04835 - Max EF 2:00:00 Trend 0.01901 - Precipitation 9:00:00 Trend 0.01706 - Avg. RF Watts 13:00:00 Trend 0.01470 - Precipitation 15:00:00 Trend

Table 5.5: A comparison of the selected features and their associated importance rankings of the top models produced by our tiered feature selection method, Recursive Feature Elimination, and Sequential Feature Selection.

Features	**R^2 Score**	**AICc Value**	**95% Confidence Interval Span**
Temperature Shortwave Radiation USU Humidity 9-hr Trend	0.45429	-15805.18	62.79
Shortwave Radiation USU Humidity 9-hr Trend	0.40935	-15582.80	65.38
Temperature Shortwave Radiation USU	0.40849	-15578.67	65.52
Shortwave Radiation USU	0.36005	-15357.48	68.12
Humidity 9-hr Trend	0.29845	-15094.56	71.24
Temperature Humidity 9-hr Trend	0.28472	-15035.08	71.92
Temperature	0.27342	-14994.26	72.46

Table 5.6: The R^2 Score, AICc value, and the non-normalized 95% confidence interval motion count span of models trained and tested using each combination of the inputs temperature, shortwave radiation, and nine-hour trend of relative humidity. A larger R^2 score, lower AICc value, and smaller confidence interval span indicates a better model.

	RFR	**PLS**	**KNN**
1st 8-Hr Avg. R^2 Score	0.77	0.76	0.71
2nd 8-Hr Avg. R^2 Score	0.73	0.82	0.53
3rd 8-Hr Avg. R^2 Score	0.66	0.72	0.40
1st 8-Hr Avg. 95% Confidence Interval Span	94.6	90.2	101.9
2nd 8-Hr Avg. 95% Confidence Interval Span	130.9	97.2	155.7
3rd 8-Hr Avg. 95% Confidence Interval Span	130.3	89.2	134.5

Table 5.7: The average R^2 score and average non-normalized 95% confidence interval motion count span of all models trained and tested for various data periodicities. The average is calculated across each 8-hour third of the 24 hour periodicities tested (i.e. 0.25 to 8, 8.25 to 16, and 16.25 to 24 hours). While the Partial Least Squares (PLS) model achieved the overall best performance across the periodicities, the Random Forest Regressor (RFR) had a slightly better R^2 score with a slightly wider confidence interval at the lower periodicities. The K-Nearest-Neighbors (KNN) model performed competitively at lower periodicities, but not as well otherwise.

CHAPTER 6

CONCLUSIONS

As part of this research effort, we have produced and made available to the public a cost effective Weather and EMR Sensing Station. The data of such a station can be paired with omnidirectional bee motion count data to train a random forest regressor model that has sufficient accuracy to estimate what the foraging activity of a bee hive should be at a given time. By making the source code and station assembly instructions available to the public, additional bee keepers may have a cost effective means of monitoring their hive's activity.

Additionally, we have produced 12 preprocessed data files containing weather and EMR data paired with bee motion count data for six different bee hives. These data sets have been made available to the public so other researchers can replicate our findings and discover new insights. Also, our hope is that researchers can use this data to have a common data set wherewith accurate predictive model comparisons can be made.

By using these new data sets, we have been able to produce new insights into how weather and ambient EMR variables correlate with bee activity, and how some of these variables can be used to predict omnidirectional bee motion. Our analysis of the data has shown that many of these variables influence bee activity in various ways, but that no individual variable can be used alone to reliably predict bee activity. For instance, while warmer temperatures may usually signify high bee activity, high wind speeds or lower shortwave radiation at the same time may decrease it. In order to achieve the highest bee motion prediction accuracies, select variables must be used jointly to produce a model that can account for different overall environmental conditions.

By generating thousands of models using different combinations of the collected variables and their engineered counterparts, we have empirically shown which combination of variables can be used to produce the most accurate bee motion predicting model. These

variables, in order from most important to least, include the nine-hour relative humidity trend, temperature, shortwave radiation, eight-hour temperature trend, relative humidity, month, pressure, five-hour shortwave radiation trend, eight-hour average wind speed trend, average wind speed, and hour.

While other research has shown that temperature and shortwave radiation are significant predictor variables of bee activity [24] [25] [26] [27], the discovery that the nine-hour relative humidity trend as a major indicator in predicting bee activity is novel as far as we are aware. When used in combination with temperature and shortwave radiation, a model is produced with an R^2 score approximately 0.05 higher than without it. Also, when the relative humidity trend is used alone, it produces a better model than when temperature is used alone. This could indicate that bees respond to how rapid and by how much the relative humidity changes over time, rather than what the relative humidity currently is.

In analyzing the effects of ambient EMR on bee hive activity at our data collection location with our selected EMR sensor, we found that it didn't significantly contribute to the predictions of omnidirectional bee traffic in comparison to the contributions made by weather variables. That said, we did observe a fairly strong inverse correlation between bee total motion at the hive entrance and the Average Total Density of the RF signals ranging from 0.01 GHz to 10 GHz, and when used alone to predict bee total motion, it was found to produce a model with an R^2 score of approximately 0.03. Additionally, when a model was trained and tested with several EMR variables together (Avg. Total Density, Avg. RF Density, Avg. RF Watts, and Avg. EF) and no weather variables, it was found that nearly 19% of the variation in the bee motion at the hive entrance could be explained by the variation of these EMR variables. Thus, EMR can be used to predict bee motion, just not with as great an accuracy as obtained by weather variables alone, nor in combination with the weather variables.

Also, while our experiments only monitored one aspect of the colony (omnidirectional motion at the hive entrance), there are many different phases of a honey bee's life that could potentially be affected by EMR that could be explored further. Additional research

is be needed in order to determine the extent of EMR's effects on bees, and whether it is contributing to their population decline. Since these experiments monitored the ambient EMR in conjunction with bee traffic at a particular location, our results could be used as a control group for future experiments with generated EMR directed at a monitored hive.

We have also shown that omnidirectional bee motion can be predicted with sufficient accuracy to produce timely bee keeper alerts with the use of on-site climate data and the random forest regressor algorithm. We've shown how our model performs at various data periodicities, and that the model can be used to meet a beekeeper's specific needs. If a beekeeper is interested in monitoring the general health of a hive, larger data prediction periodicities (such as 12 hours) can be used with exceptionally high levels of accuracy (such as an R^2 score of 0.89). This may be sufficient to alert if a hive is suffering from something like Varroa mites or other long-term hive degrading conditions. On the other hand, if a beekeeper is more interested in being alerted of a swarming, robbing, or other immediately impactful hive event, a much shorter periodicity (such as 15 minutes to 1 hour) can be used instead with lower accuracies (such as R^2 scores of 0.54 to 0.72). In the case of these smaller periodicities, it may be possible decrease false positives by utilizing a threshold, such as the 95% confidence interval span, so alerts are only triggered when the observed values are beyond the predicted values plus or minus the threshold. Such alerts would allow beekeepers to effectively and efficiently monitor their hives from a distance, and be able to know when specific hives need attention.

Our hive entrance bee activity prediction accuracies are on-par with other hive entrance bee activity prediction accuracies reported in the literature. As mentioned before, since the studies found by the authors were performed in different locations and used different data sets, the results can't be accurately ranked. Regardless, at one hour periodicities, one study performed by Clarke and Robert in the United Kingdom achieved R^2 scores between 0.79 and 0.81 [24], and another study performed by Devillers et al. in France achieved R^2 scores between 0.62 and 0.72 [26]. For reference, we achieved R^2 scores of 0.72 for both the R_4_5 and R_4_11 hives at one hour periodicities.

While the Clarke and Robert study achieved higher R^2 scores, as we noted before, their data exhibited a much higher collective average correlation strength across their variables that overlapped with ours. Specifically, their average Pearson correlation strength over the magnitudes of shortwave radiation, temperature, and humidity was 0.75 while ours was 0.37. These differences could be attributed to any number of differences including location, data collection methods, the weather of the data collection season, or even the phenology of specific bee colonies. Devillers et al. didn't report the correlation coefficients for their collected variables. If the data used in our experiments had the same correlation coefficients as the Clarke and Robert study, we would anticipate that our model's prediction accuracies would be much higher. Unfortunately, we could not find Clarke and Robert's data for experimentation and comparison.

While both of the above-mentioned studies used PLS for predicting bee activity, we showed that a random forest regressor can perform competitively with the PLS algorithm, and better than the KNN algorithm for this predictive problem. Additionally, it lends itself well to giving insight into how important each feature is in making predictions of bee total motion from one moment to the next. However, while these feature importances don't necessarily reveal to what the bees are actually be responding, they may give valuable insight to entomologists and lead to significant discoveries.

As mentioned in the introduction, our research aimed to conform to several guiding principles. One of these included that the bee space be preserved such that the EBM sensors don't disrupt any natural bee colony cycles. We were able to maintain this principle by utilizing the BeePi monitoring system (which uses external video data to estimate the hive traffic) [20], and by collecting external environmental metrics for making predictions. The other research discussed in this thesis utilized sensors placed in the hive or entrance to collect some of the data used to make predictions. While these means may have enabled more accurate data collection processes and thereby more accurate predictions, it comes at the cost of interfering with the natural bee hive cycles. Additionally, it could discourage bee keepers from adopting such methods of bee hive monitoring. We've shown that comparable

bee activity prediction accuracies can be achieved without use of invasive data collection means.

As bee populations decline, our predictive model, made possible with cost effective equipment, could be a valuable asset to apiarists around the world in identifying common trends that could lead to important discoveries. With timely alerting possible, apiarists can discover hive issues before they become too late to resolve. If data collected by hive monitors around the world is made available to the public, a more complete picture could be formed of honey bee populations at large. This could lead to more advanced remedies to combat the forces causing bee decline, more cost effective hive maintenance practices, and increased insight into honey bee behavior. Additionally, animals and insects have often been used as alerting mechanisms for potential problems to humans. If the decline in bee populations happens to be a harbinger of issues relevant to us, it's in our best interest to understand what may be causing it, and do our part to resolve them before it's too late.

CHAPTER 7

FUTURE WORK

The scope of this research leaves many facets available for future exploration and improvement. Firstly, our Weather and EMR Sensing station could be enhanced. Since these experiments utilized the first iteration of the station, limited prototype testing was performed before use. This led to the exposure of bugs and hardware failures while the system was operational at the site. Some of these included water shorting the power supply, condensation dripping into the Raspberry Pi's SD Card slot and shorting it, the Pi not utilizing the Real Time Clock properly allowing the time to get off after a power outage, and crashes observed when the temperature fell below 26°F.

While we were able to resolve these bugs and issues as they were encountered, each hardware failure resulted in a gap in the data. Fortunately, these issues were minor and caused little down time. However, the station should be made more robust so fewer data collection gaps occur, and further testing should be performed to vet out any other issues that may present themselves in different situations.

Additionally, it may be beneficial to utilize a more accurate EMR sensor. In an effort to keep the cost of building the station minimal, we chose to use the EMF-390 sensor. While this came with the benefit of having many sensors integrated into a hand-held unit and its ability to integrate well with a Raspberry Pi for long-term logging, it's possible that a higher-end sensor could detect things that this sensor could not. As mentioned earlier, we discovered that temperature changes interfered with the EMF variable's readings. This forced us to remove this variable entirely from the dataset. Perhaps if a more consistently accurate sensor was used, further insights could be obtained into how EMF affects bees. Also, if one monitored variable was a little skewed, it could be possible that other environmental factors could be skewing the readings as well. While we took every effort to detect any incorrect readings, it's possible that some were missed which could have impacted the

results. Additional testing and sensor exploration is needed.

It should also be noted that all of our data logging with the EMF-390 sensor was performed with the sensor mounted vertically with the screen facing north. The documentation for the device indicates that the EMF sensing component can monitor on the x, y, and z axes, however, it also states that for the most accurate RF readings, the top of the device should be pointed toward a desired RF source to monitor. Since we were focused on ambient EMR and weren't monitoring any particular RF source, we chose to mount the device vertically. Nevertheless, additional data could be collected with the sensor pointed in different directions to get more directional RF readings at the data collection location. For instance, the sensor could be pointed toward a radio or cell tower so RF from specific sources could be correlated with bee activity.

On that note, since our research encompassed the effects of ambient EMR at the hive, and we have recorded its predictive ability on bee total motion at the hive entrance, additional experiments could be performed where generated EMR is produced and directed at the hive. This would allow bee activity to be monitored in a more EMR-controlled setting.

Another aspect of the EMR sensor that could be leveraged, is the monitoring of different frequency bands. While the Avg. Total Density variable collects readings from all frequencies from 0.01 GHz to 10 GHz, the other RF variables we collected, such as Avg. RF Watts, were more specifically configured to monitor frequencies in greater detail between 240 MHz and about 1040 MHz. This was due to the device only being able to monitor one specific band at a time for those particular variables collected. However, the device can be configured to monitor four other frequency bands: 50 MHz - 65 MHz, 65 MHz - 76 MHz, 76 MHz - 108 MHz, and 2.4 GHz - 2.5 GHz. Additional experiments could be performed while the device is reconfigured to monitor each of the different frequency bands in turn to see if the bees are affected by one band more than another.

Also, since our research encompassed only a single season, collecting additional data would likely improve our models. This would allow further training of the random forest

regressor model so it could better recognize patterns between bee activity and environmental factors. Since weather patterns can vary quite a bit from one season to another, our model is currently tuned to the particular weather patterns that were experienced during our data collection season. Additional data would help the model generalize better from one year to the next and allow for better testing and model refinement.

More data could also lead to new discoveries regarding bee activity and environmental factors. For instance, in our research we noticed some interesting correlations that could be explored further. This could include the effects of wind speed on bee activity. While we observed a generally positive correlation between bee traffic and wind speed, further investigation revealed that from one hour to the next, there may actually be a negative correlation. It may be that there is a threshold at which the wind speed begins to affect bee traffic. We also observed some instances where bee activity was much higher than normal while the pressure steeply dropped before precipitation was received. Perhaps bees can somewhat "predict" precipitation. While we observed these phenomena on some occasions, we did not collect sufficient data to make any conclusions, and more research would be needed to explore these areas further.

It would also be desirable to collect data from different locations. All of our data was collected not far from the city center in a residential area of Logan, Utah. Further insight into how bees behave within urban areas compared to bees in rural areas could be obtained by collecting data on hives outside of the city for comparison. It's likely that the EMR levels and weather factors would be different outside the city.

While we collected many different weather and EMR variables, it would have also been valuable to monitor the absolute humidity. Since we discovered that the relative humidity trend was a major predictor in bee activity, it could be possible that the absolute humidity trend could be an even better predictor of bee activity. Other variables would have been valuable to explore as well such as the dew point, CO_2 levels, or other solar radiation variables.

In all of our research, we only monitored the omnidirectional bee traffic. While this

gives us insight into one aspect of the hive activity, the honey bee life cycle encompasses many other facets. It may be possible that higher EMR values or different weather conditions don't affect bee motion at the hive entrance, but they could influence other bee activity inside the hive or while foraging. Other response variables could be analyzed, such as hive buzz intensity, to see if there are any connections with changes in the environment.

Additionally, since we observed a much lower overall correlation between some of our collected variables and bee motion than other research [24], it's possible that this is indicative that our bee motion counting technology could be further enhanced. With more accurate data being fed into the prediction model, more accurate predictions will likely be produced by it.

Lastly, while we explored several different predictor models including a Random Forest Regressor, Partial Least Squares Regression, and K-Nearest-Neighbors, many other models could be explored as well. This could include Time Series, Neural Networks, or Reinforcement Learning algorithms. While we attempted to incorporate a temporal presence in our data by adding a month and hour column, an algorithm more suited for time series analysis could potentially produce a more accurate model. Also, since we discovered that no one variable alone could predict the bee motion with a high level of accuracy, it could be beneficial to utilize a neural network. A model such as this could potentially be trained to detect more complex relationships that the conceptually simple random forest model could not. Of course, these models would have to be evaluated not only on the accuracy they can achieve, but the practicality of their use in bee hive monitoring situation on hardware with limited resources.

Our hope is that the research described in this thesis can open the doors to many other new and exciting discoveries. Bees are very complex insects with sophisticated life cycles, daily activities, and yearly patterns. As automated hive monitoring and alerting is improved, we will be able to help bee keepers become more efficient in their labors and increase the information we have to help slow the decline of bee populations.

REFERENCES

[1] U.S. Food & Drug Administration, "Helping agriculture's helpful honey bees," September 2021. [Online]. Available: https://www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-honey-bees

[2] U.S. Department of Agriculture, "Us pollinator information," September 2021. [Online]. Available: https://www.ree.usda.gov/pollinators

[3] National Agricultural Statistics Service, "Honey bees," National Agricultural Statistics Service, Tech. Rep., September 2019. [Online]. Available: https://www.nass.usda.gov/Publications/Highlights/2019/2019_Honey_Bees_StatisticalSummary.pdf

[4] National Agricultural Statistics Service, Agricultural Statistics Board, and U.S. Department of Agriculture, "Honey bee colonies," National Agricultural Statistics Service, Tech. Rep., August 2021. [Online]. Available: https://downloads.usda.library.cornell.edu/usda-esmis/files/rn301137d/8g84nk42x/00000x890/hcny0821.pdf

[5] D. vanEngelsdorp and M. D. Meixner, "A historical review of managed honey bee populations in europe and the united states and the factors that may affect them," *Journal of Invertebrate Pathology*, vol. 103, pp. S80–S95, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022201109001827

[6] U.S. Department of Agriculture, "Colony collapse disorder action plan," June 2007. [Online]. Available: https://www.ars.usda.gov/is/br/ccd/ccd_actionplan.pdf

[7] P. J. Caradonna, J. L. Cunningham, and A. M. Iler, "Experimental warming in the field delays phenology and reduces body mass, fat content and survival: implications for the persistence of a pollinator under climate change," *Functional Ecology*, vol. 32, pp. 2345–2356, 2018. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2435.13151

[8] B. Cornelissen, P. Neumann, and O. Schweiger, "Global warming promotes biological invasion of a honey bee pest," *Global Change Biology*, vol. 25, pp. 3642–3655, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31394018/

[9] R. D. Girling, I. Lusebrink, E. Farthing, T. A. Newman, and G. M. Poppy, "Diesel exhaust rapidly degrades floral odours used by honeybees," *Scientific Reports*, vol. 3, 2013. [Online]. Available: https://doi.org/10.1038/srep02779

[10] S. Shepherd, G. Hollands, V. C. Godley, S. M. Sharkh, C. W. Jackson, and P. L. Newland, "Increased aggression and reduced aversive learning in honey bees exposed to extremely low frequency electromagnetic fields," *PLoS ONE*, vol. 14, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0223614

[11] B. Greenberg, V. P. Bindokas, M. J. Frazier, and J. R. Gauger, "Response of honey bees, apis mellifera l., to high-voltage transmission lines," *Environmental Entomology*, vol. 10, pp. 600–610, 1981. [Online]. Available: https://doi.org/10.1093/ee/10.5.600

[12] S. Shepherd, G. Hollands, V. C. Godley, S. M. Sharkh, C. W. Jackson, and P. L. Newland, "Increased aggression and reduced aversive learning in honey bees exposed to extremely low frequency electromagnetic fields," *PLoS ONE*, vol. 14, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0223614

[13] K. E. Smith, D. Weis, M. Amini, A. E. Shiel, V. W.-M. Lai, and K. Gordon, "Honey as a biomonitor for a changing world," *Nature Sustainability*, vol. 2, pp. 223–232, 2019. [Online]. Available: https://www.nature.com/articles/s41893-019-0243-0

[14] K. Bunzl, W. Kracke, and G. Vorwohl, "Transfer of chernobyl-derived 134cs, 137cs, 131i and 103ru from flowers to honey and pollen," *Journal of Environmental Radioactivity*, vol. 6, no. 3, pp. 261–269, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0265931X88900811

[15] I. C. Elżbieta Skorbiłowicz, Mirosław Skorbiłowicz, "Bees as bioindicators of environmental pollution with metals in an urban area," *Journal of Ecological Engineering*, vol. 19, no. 3, pp. 229–234, May 2018. [Online]. Available: https://doi.org/10.12911/22998993/85738

[16] J.-A. Jiang, C.-H. Wang, C.-H. Chen, M.-S. Liao, Y.-L. Su, W.-S. Chen, C.-P. Huang, E.-C. Yang, and C.-L. Chuang, "A wsn-based automatic monitoring system for the foraging behavior of honey bees and environmental factors of beehives," *Computers and Electronics in Agriculture*, vol. 123, pp. 304–318, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169916300709

[17] X.-J. He, Liu-Qing, X.-B. Wu, and Z.-J. Zeng, "Rfid monitoring indicates honeybees work harder before a rainy day," *Insect Science*, vol. 23, pp. 157–159, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/1744-7917.12298

[18] V. Kulyukin, S. Mukherjee, and P. Amlathe, "Toward audio beehive monitoring: Deep learning vs. standard machine learning in classifying beehive audio samples," *Applied Sciences*, vol. 8, no. 9, 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/9/1573

[19] W. G. Meikle, N. Holst, T. Colin, M. Weiss, M. J. Carroll, Q. S. McFrederick, and A. B. Barron, "Using within-day hive weight changes to measure environmental effects on honey bee colonies," *PLoS ONE*, vol. 13, no. 5, 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0197589

[20] V. Kulyukin and S. Mukherjee, "On video analysis of omnidirectional bee traffic: Counting bee motions with motion detection and image classification," *Applied Sciences*, vol. 9, no. 18, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/18/3743

[21] V. Kulyukin, S. Mukherjee, A. Minichiello, and T. Truscott, "Beepiv: A method to measure apis mellifera traffic with particle image velocimetry in videos," *Applied Sciences*, vol. 11, no. 2276, 2021. [Online]. Available: https://doi.org/10.3390/app11052276

[22] A. R. Braga, D. G. Gomes, R. Rogers, E. E. Hassler, B. M. Freitas, and J. A. Cazier, "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies," *Computers and Electronics in Agriculture*, vol. 169, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919317661?via%3Dihub

[23] R. M. Burrill and A. Dietz, "The response of honey bees to variations in solar radiation and temperature," *Apidologie*, vol. 12, no. 4, pp. 319–328, 1981. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00890551/document

[24] D. Clarke and D. Robert, "Predictive modelling of honey bee foraging activity using local weather conditions," *Apidologie*, vol. 49, pp. 386–396, 2018. [Online]. Available: https://doi.org/10.1007/s13592-018-0565-3

[25] L. P. Polatto, J. Chaud-Netto, and V. V. Alves-Junior, "Influence of abiotic factors and floral resource availability on daily foraging activity of bees," *Journal of Insect Behavior*, vol. 27, pp. 593–612, 2014. [Online]. Available: https://link.springer.com/article/10.1007/s10905-014-9452-6

[26] J. Devillers, J. Dore, M. Tisseur, S. Cluzeau, and G. Maurin, "Modelling the flight activity of apis mellifera at the hive entrance," *Computers and Electronics in Agriculture*, vol. 42, no. 2, pp. 87–109, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169903001029

[27] I. M. de Mattos, J. Souza, and E. Soares, "Analysis of the effects of climate variables on apis mellifera pollen foraging performance," *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, vol. 70, pp. 1301–1308, 08 2018.

[28] A. Zacepins, A. Kviesis, E. Stalidzans, M. Liepniece, and J. Meitalovs, "Remote detection of the swarming of honey bee colonies by single-point temperature monitoring," *Biosystems Engineering*, vol. 148, pp. 76–80, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1537511016300964

[29] L. H. S. Alves, P. C. R. Cassino, and F. Prezoto, "Effects of abiotic factors on the foraging activity of apis mellifera linnaeus, 1758 in inflorescences of vernonia polyanthes less (asteraceae)," *Acta Scientiarum. Animal Sciences*, vol. 37, p. 405, 10 2015.

[30] G. Hennessy, C. Harris, C. Eaton, P. Wright, E. Jackson, D. Goulson, and F. F. Ratnieks, "Gone with the wind: effects of wind on honey bee visit rate and foraging behaviour," *Animal Behaviour*, vol. 161, pp. 23–31, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003347220300026

[31] H. Xujiang, L.-Q. Tian, X. Wu, and Z.-J. Zeng, "Rfid monitoring indicates honeybees work harder before a rainy day," *Insect science*, vol. 23, 11 2015.

[32] S. Kimmel, J. Kuhn, W. Harst, and H. Stever, "Electromagnetic radiation: Influences on honeybees (apis mellifera). iias-intersymp conference," *Baden-Baden*, pp. 1–6, 01 2007.

[33] R. R. Taye, M. K. Deka, A. Rahman, and M. Bathari, "Effect of electromagnetic radiation of cell phone tower on foraging behaviour of asiatic honey bee, apis cerana f. (hymenoptera: Apidae)," *Journal of entomology and zoology studies*, vol. 5, pp. 1527–1529, 2017.

[34] Raspberry Pi Foundation, "Build your own weather station," June 2021. [Online]. Available: https://projects.raspberrypi.org/en/projects/build-your-own-weather-station

[35] Bosch. (2020) Bme280 combined humidity and pressure sensor. [Online]. Available: https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme280-ds002.pdf

[36] R. Hull, "Rpi.bme280 0.2.3," June 2021. [Online]. Available: https://pypi.org/project/RPi.bme280/

[37] Argent Data Systems, "Wind / rain sensor assembly," June 2021. [Online]. Available: https://www.argentdata.com/catalog/product_info.php?products_id=145

[38] ——. (2021) Weather sensor assembly p/n 80422. [Online]. Available: https://www.argentdata.com/files/80422_datasheet.pdf

[39] Apogee Instruments, "Sp-110-ss: Self-powered pyranometer," June 2021. [Online]. Available: https://www.apogeeinstruments.com/sp-110-ss-self-powered-pyranometer/

[40] ——. (2020) Owner's manual. [Online]. Available: https://www.apogeeinstruments.com/content/SP-110-manual.pdf

[41] GQ Electronics, "Advanced gq emf-390 multi-field, multi-function emf meter and rf spectrum power analyzer," June 2021. [Online]. Available: https://www.gqelectronicsllc.com/comersus/store/comersus_viewItem.asp?idProduct=5678

[42] Apogee Instruments. (2020) Owner's manual. [Online]. Available: https://www.apogeeinstruments.com/content/SP-110-manual.pdf

[43] D. D. Farra, "em390cli," June 2021. [Online]. Available: https://gitlab.com/codref/em390cli

[44] Raspberry Pi Foundation. (2021) Raspberry pi 3 model b+. [Online]. Available: https://datasheets.raspberrypi.org/rpi3/raspberry-pi-3-b-plus-product-brief.pdf

[45] Microchip. (2008) Mcp3204/3208. [Online]. Available: https://ww1.microchip.com/downloads/en/DeviceDoc/21298e.pdf

[46] macetech, "Chronodot," June 2021. [Online]. Available: https://docs.macetech.com/doku.php/chronodot_v2.0

[47] B. Nuttall, "gpiozero 1.6.2," July 2021. [Online]. Available: https://pypi.org/project/gpiozero/

[48] Federal Communications Commission, "Rf safety faq," July 2021. [Online]. Available: https://www.fcc.gov/engineering-technology/electromagnetic-compatibility-division/radio-frequency-safety/faq/rf-safety

[49] U.S. Department of Commerce, "United states frequency allocations," January 2016. [Online]. Available: https://www.ntia.doc.gov/files/ntia/publications/january_2016_spectrum_wall_chart.pdf

[50] S. Mukherjee and V. Kulyukin, "Application of digital particle image velocimetry to insect motion: Measurement of incoming, outgoing, and lateral honeybee traffic," *Applied Sciences*, vol. 10, no. 2042, 2020. [Online]. Available: https://doi.org/10.3390/app10062042

[51] Utah Climate Center, "Usu environmental observatory," September 2020. [Online]. Available: https://climate.usu.edu/mchd/dashboard/dashboard.php?network=USUwx&station=1279257&units=E&showgraph=0&

[52] ——, "Usu weather," September 2020. [Online]. Available: https://climate.usu.edu/mchd/dashboard/overview/USUwx.php

[53] National Weather Service, "Air pressure," August 2021. [Online]. Available: https://www.weather.gov/jetstream/pressure

[54] ——, "Relative humidity," July 2021. [Online]. Available: https://w1.weather.gov/glossary/index.php?word=relative+humidity

[55] ——, "Jetstream max: The ionosphere," August 2021. [Online]. Available: https://www.weather.gov/jetstream/ionosphere_max

[56] Federal Communications Commission, "Why am stations must reduce power, change operations, or cease broadcasting at night," August 2021. [Online]. Available: https://www.fcc.gov/media/radio/am-stations-at-night

[57] A. Felix, A. A. Olufemi, H. D. Ibrahim, A. Ayegba, J. J. Olu, W. D. Fonyuy, and A. Victor, "Investigation of the influence of atmospheric temperature and relative humidity on fm radio signal strength: A case study of we fm abuja," *International Journal Of Scientific & Technology Research*, vol. 6, no. 11, pp. 70–74, 2017. [Online]. Available: https://www.semanticscholar.org/paper/Investigation-Of-The-Influence-Of-Atmospheric-And-A-Felix-Olufemi/def39410ca8e17786906959b5f981ccd06e5e6ce

[58] S. A. Mirbagheri and M. Mohammadi, "Prediction of environmental effects in received signal strength in fm/tv station based on meteorological parameters using artificial neural network and data mining," *Journal of Environmental Management*, vol. 250, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301479719311727

[59] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[61] scikit-learn, "Ensemble methods," August 2021. [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html#forest

[62] ——, "Decision trees," August 2021. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html#decision-trees

[63] ——, "Decision trees," August 2021. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score-the-coefficient-of-determination

[64] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974. [Online]. Available: https://ieeexplore.ieee.org/document/1100705

[65] scikit-learn, "Feature selection," September 2021. [Online]. Available: https://scikit-learn.org/stable/modules/feature_selection.html#rfe

[66] ——, "Feature selection," September 2021. [Online]. Available: https://scikit-learn.org/stable/modules/feature_selection.html#sequential-feature-selection

[67] ——, "Plsregression," August 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html?highlight=pls#sklearn.cross_decomposition.PLSRegression

[68] ——, "Kneighborsregressor," August 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

APPENDIX

This section contains some additional plots and figures that illustrate some observed connections between bee motion and environmental factors. Also included are some noteworthy observations between various weather and EMR variables.

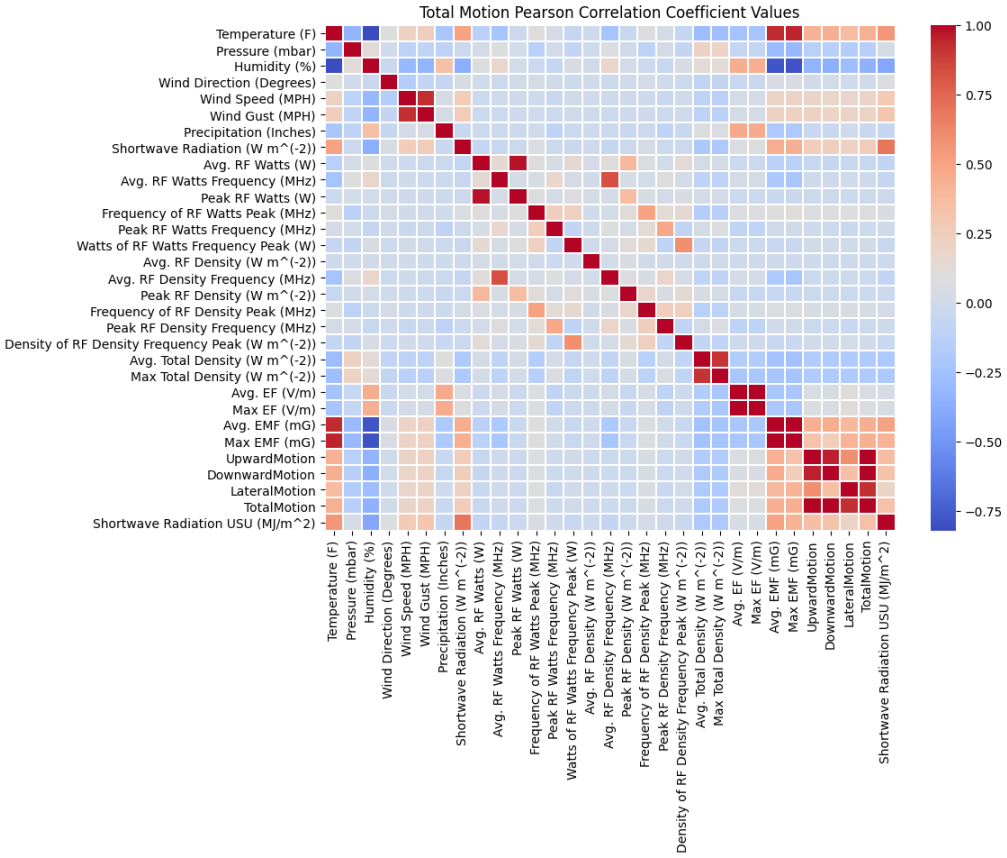

Fig. 1: A heat map showing Pearson's correlation coefficient of each column with every other column. Positive correlations are represented by red hues, and negative correlations are represented by blue hues.

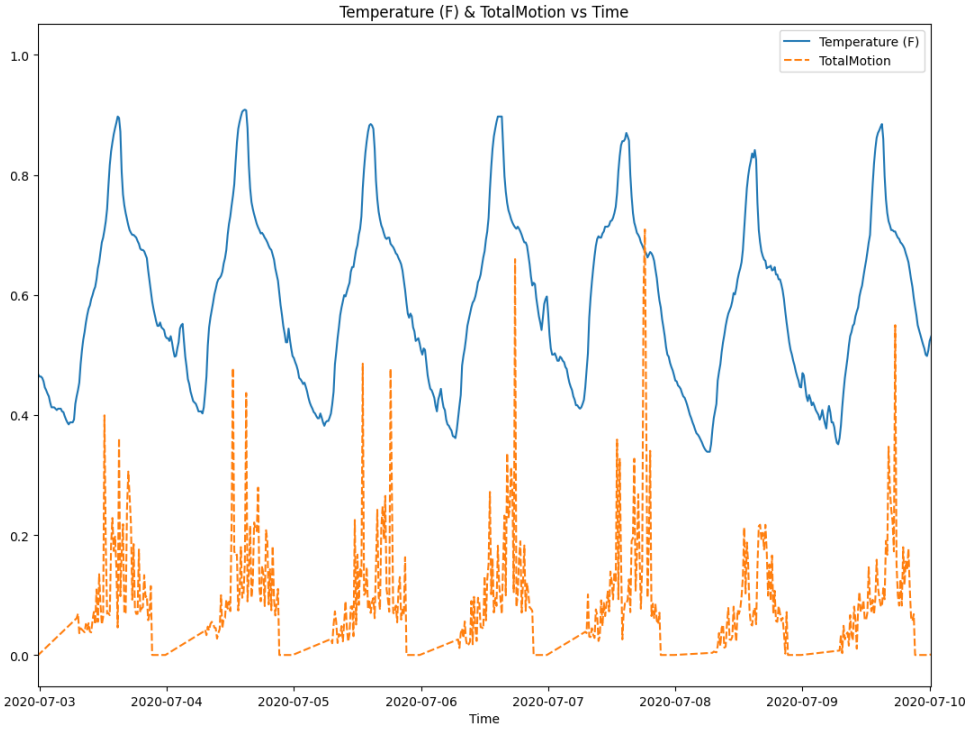

Fig. 2: Normalized Temperature and Total Motion plotted against time. Note that the peak bee motion is typically before and after the peak temperature.

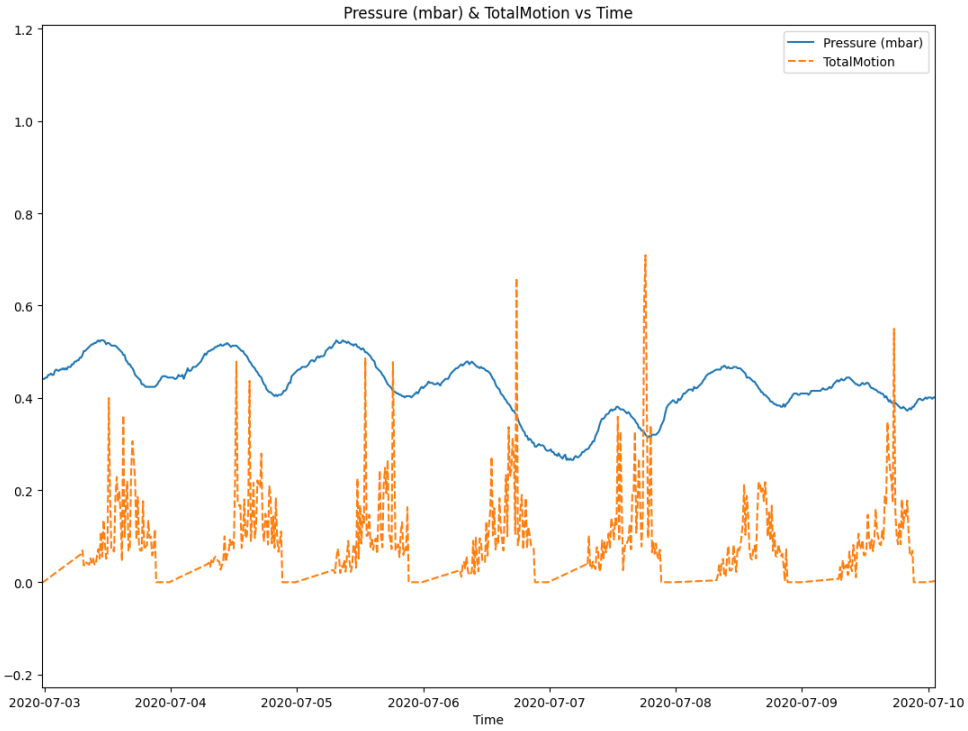

Fig. 3: Normalized Atmospheric Pressure and Total Motion plotted against time. Note that the atmospheric pressure follows a diurnal cycle, and that bee motion often peaks while the pressure is falling.

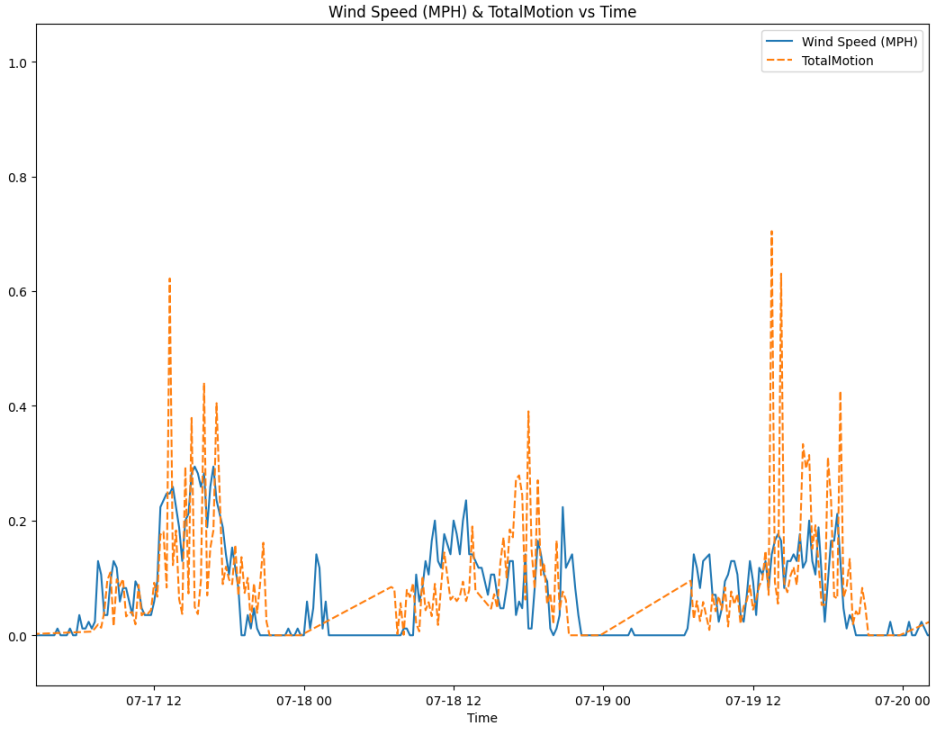

Fig. 4: Normalized Average Wind Speed and Total Motion plotted against time over a few days. This shows the generally positive correlation between the variables. Note that the wind is generally higher during the day than at night, which coincides with the diurnal activity of the bees.

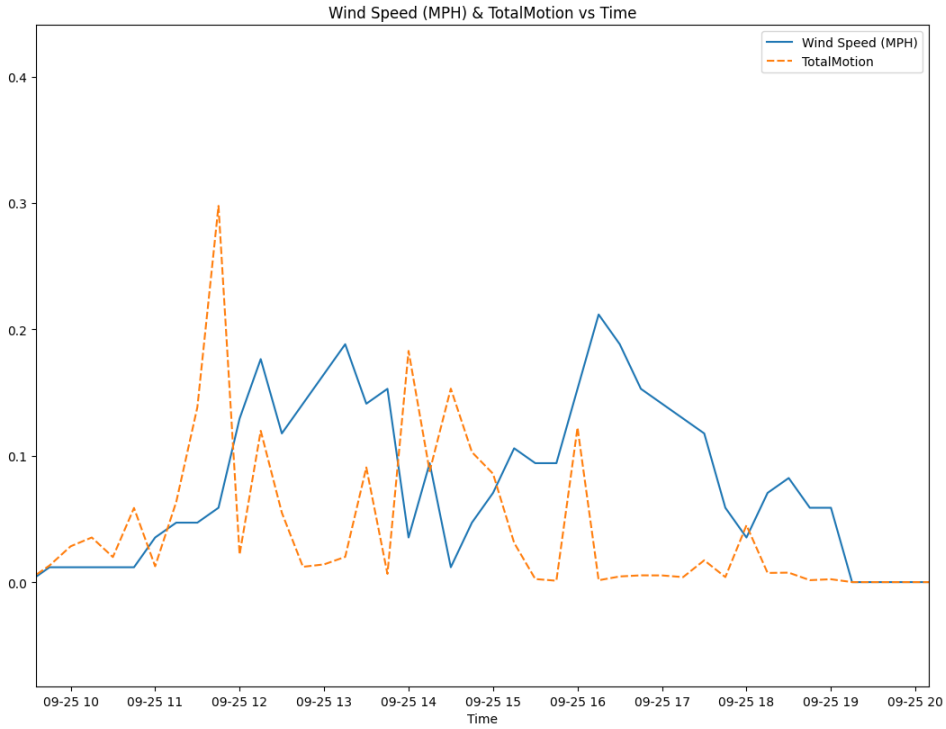

Fig. 5: Normalized Average Wind Speed and Total Motion plotted against time. Note that although the bee activity and the wind speed follow the same general daily pattern, from one hour to the next the bee motion sometimes actually drops or ceases when there are spikes in wind speed. This may indicate a localized inverse relationship between wind speed and bee motion.

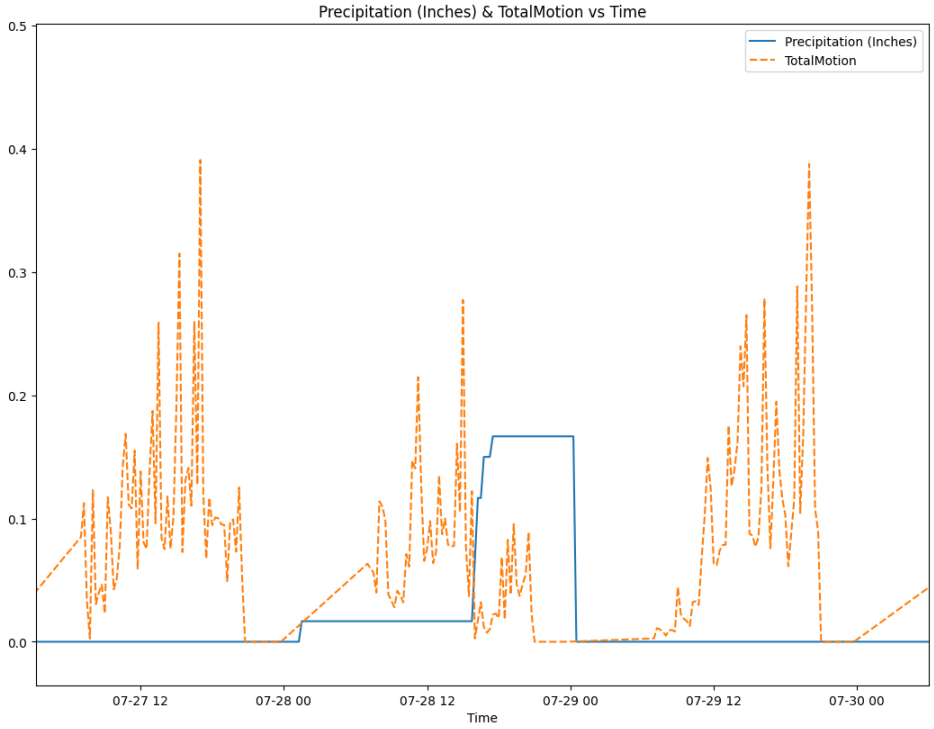

Fig. 6: Normalized Precipitation and Total Motion plotted against time. Note that bee total motion drops significantly while rain is falling (indicated by the increasing precipitation values). Also, note that bee motion resumes once the rain stops (as indicated by the flat-lined precipitation values).

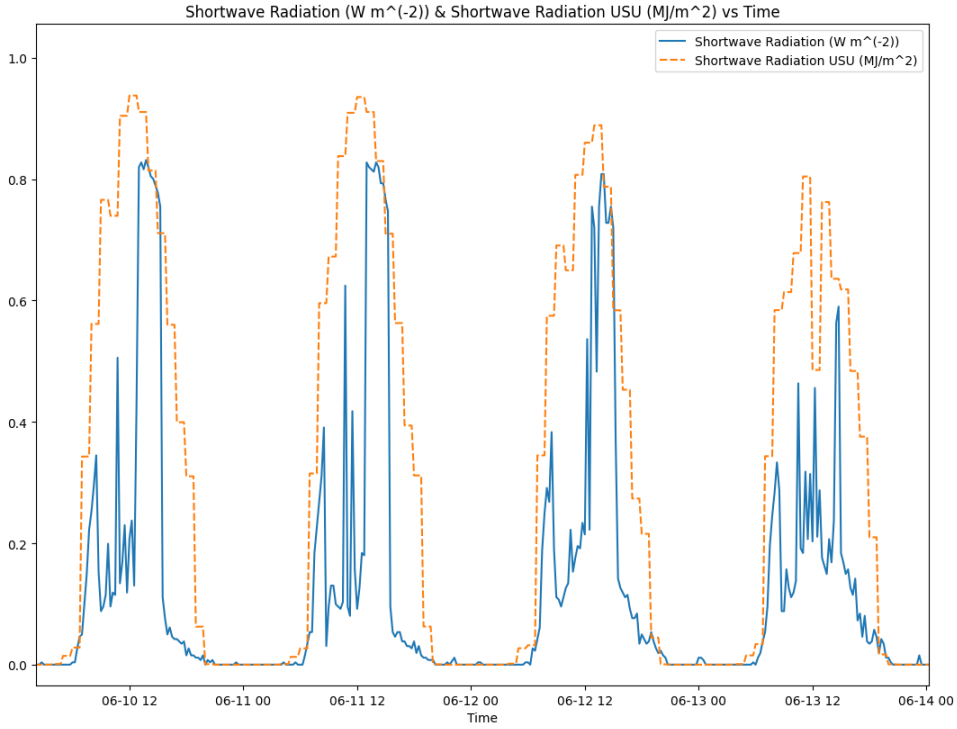

Fig. 7: Normalized Shortwave Radiation data collected from our Weather and EMR Sensing Station and the Utah Climate Center's station plotted against time. Note the irregular pattern of the data collected by our station. This was due to the foliage surrounding the location of the bee hives casting shadows on the pyranometer throughout the day.

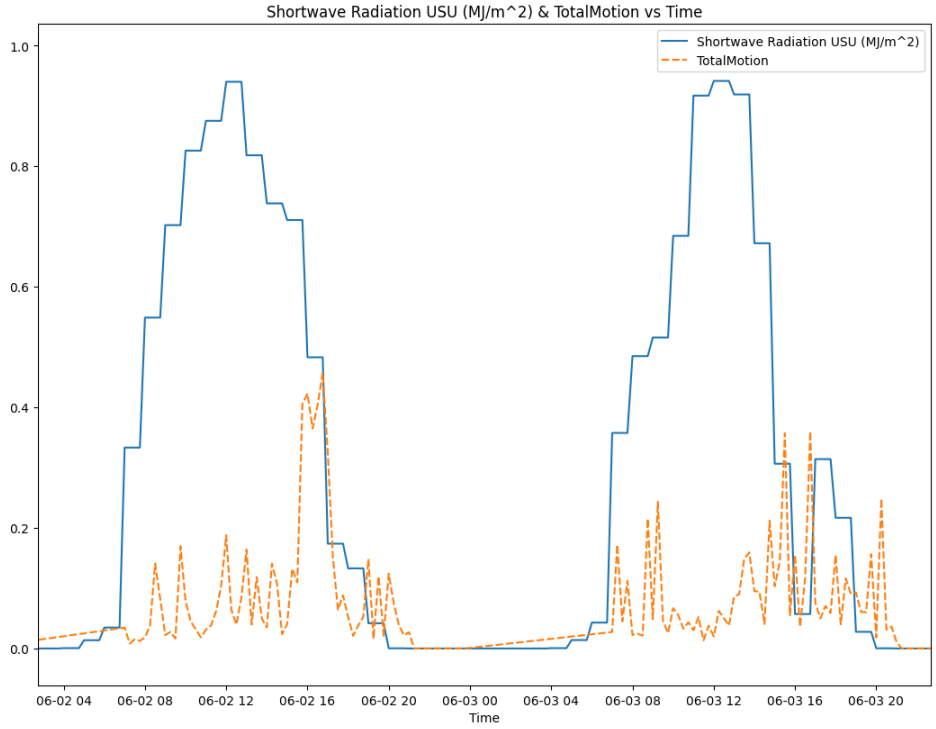

Fig. 8: Normalized Shortwave Radiation and Total Motion plotted against time. Note the drop in both Shortwave Radiation and bee Total Motion on the 3rd at the same time the peak bee activity occurred on the 2nd. This is indicative of the strong positive correlation between the two variables.

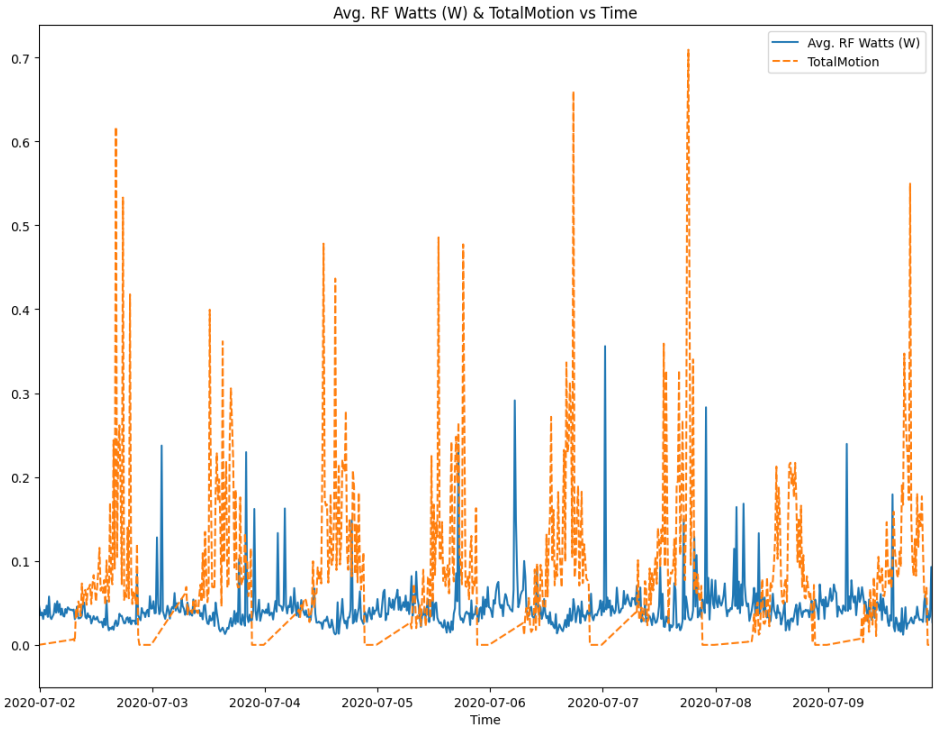

Fig. 9: Normalized Avg. RF Watts and Total Motion plotted against time. Note that the RF Watts value rises at night and falls during the day, exhibiting a negative correlation with bee total motion. This is likely due to the ionosphere reflecting more signals at night.

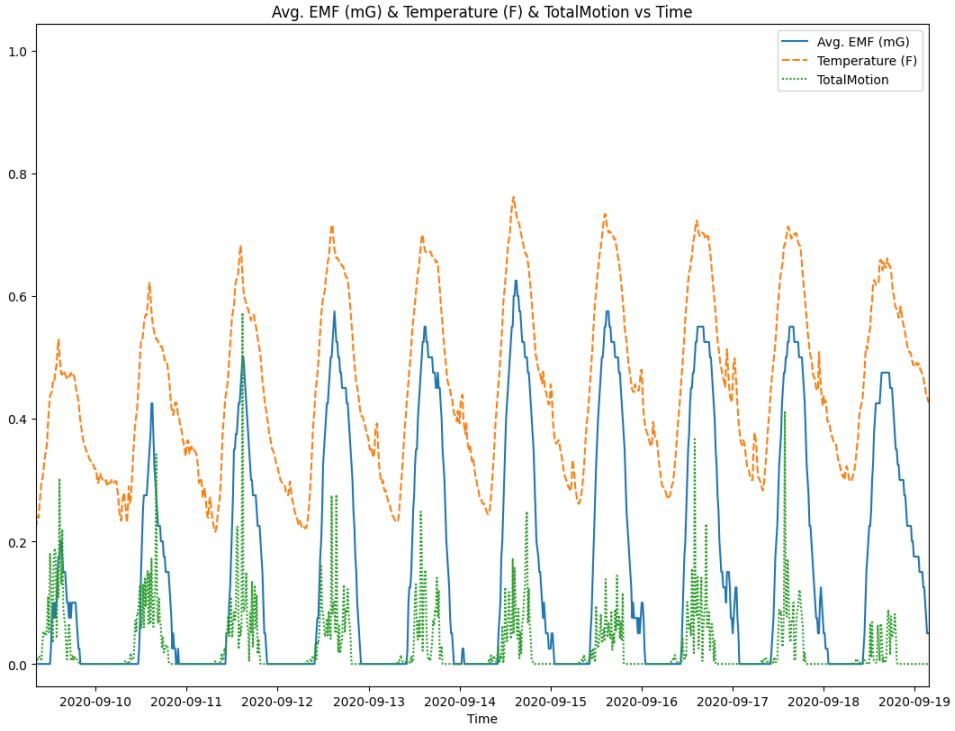

Fig. 10: Normalized Avg. EMF, Temperature, and Total Motion plotted against time. Note that the EMF values very closely follow the temperature patterns.

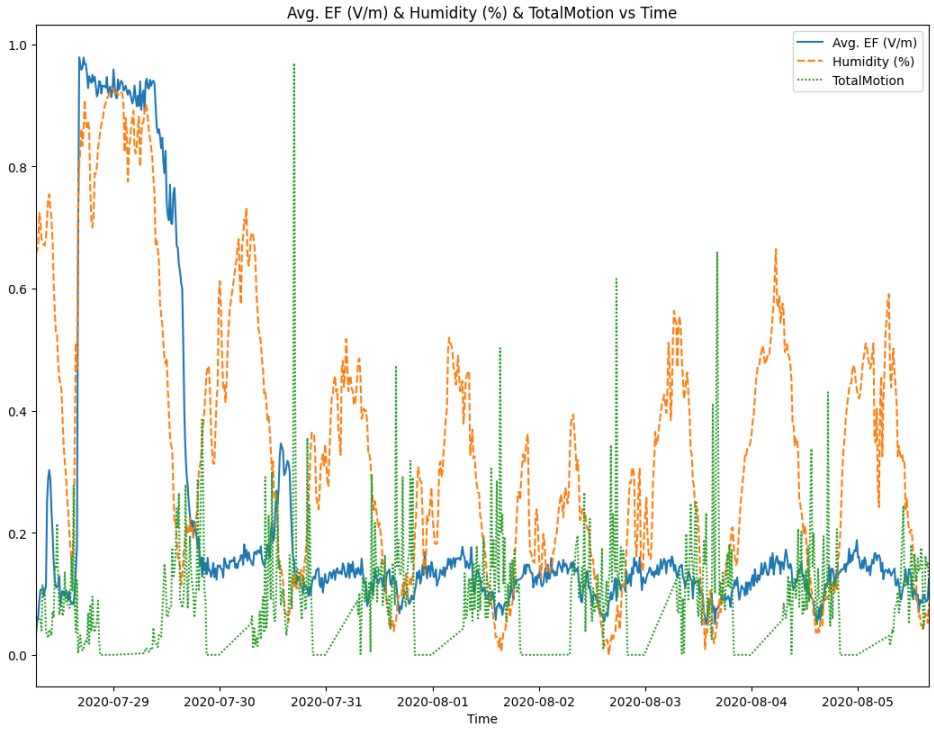

Fig. 11: Normalized Avg. EF, Relative Humidity, and Total Motion plotted against time. Note that the EF values typically rise with the relative humidity values, and that an inverse relationship between humidity and total motion appears to exist. The spike in EF on July 29th coincides with precipitation that was recorded that day, which elevated the relative humidity.